DOCTORAL THESIS

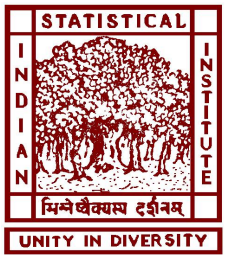
Bayesian joint modeling of multivariate longitudinal and event-time outcomes with applications to ALL maintenance studies

Author:
Damitri Kundu

Supervisor:
Dr. Kiranmoy Das

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the



Indian Statistical Institute
203, B.T. Road
Kolkata -700108.

June, 2023

Declaration of Authorship

I, Damitri Kundu, declare that this thesis titled “Bayesian joint modeling of multivariate longitudinal and event-time outcomes with applications to ALL maintenance studies” and the works presented in it are my own. I confirm the following statements.

- The work was done while in candidature for a research degree at Indian Statistical Institute, Kolkata.
- No part of this thesis has been submitted for a degree or any other qualification at this institute or elsewhere.
- I have clearly cited all the previously published works that I have consulted for writing this thesis.
- I have properly acknowledged all main sources of data used in this thesis and the methods on which I have built my work.
- I have properly acknowledged all my co-authors who helped in developing different parts of this thesis.

Signed:

_____

Date:

_____

Certificate from Supervisor

This is to certify that the thesis titled "Bayesian joint modeling of multivariate longitudinal and event-time outcomes with applications to ALL maintenance studies" submitted by ***Ms. Damitri Kundu****, who got her name registered on* ***14.07.2018*** *for the award of Ph.D. (Statistics) degree from Indian Statistical Institute, Kolkata is absolutely based upon her own work under the supervision of* ***Dr. Kiranmoy Das*** *and that neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award elsewhere before.*

Signature of the Supervisor
and date with official seal

List of Publications

Published Articles

[1] **D. Kundu**, P. Sarkar, M. Gogoi and K. Das. (2023). "A Bayesian joint model for multivariate longitudinal and time-to-event data with application to ALL maintenance studies", *Journal of Biopharmaceutical Statistics*, doi: 10.1080/10543406.2023.2187413.

[2] P. Kedia, **D. Kundu** and K. Das. (2022). "A Bayesian variable selection approach to longitudinal quantile regression", *Statistical Methods & Applications*, 32, 149–168.

[3] S. Sen, **D. Kundu** and K. Das. (2022). "Variable selection for categorical response: a comparative study", *Computational Statistics*, 38, 809–826.

[4] S. Das, **D. Kundu** and A. Dewanji. (2022). "Software reliability modeling based on NHPP for error occurrence in each fault with periodic debugging schedule", *Communications in Statistics - Theory and Methods*, 51, 4890–4902.

Articles under revision/ review

[1] **D. Kundu**, K. Das. (2023). "A Bayesian quantile joint modeling for multivariate longitudinal and event-time data", *revision invited from Lifetime Data Analysis*

[2] S. Sen, **D. Kundu**, K. Das. (2023). "A flexible Bayesian approach for modelling interval data", *revision invited from Statistical Methods & Applications*

[3] **D. Kundu**, K. Das. (2023). "A Bayesian latent class joint model for multivariate longitudinal and time-to-event data", *under review*

Abstract

Joint analysis of longitudinal and event-time outcomes is a major research topic in the last two decades, mainly due to its successful applications in various disciplines including medical studies, biological studies, environmental studies, economics and many others. When a group of individuals are followed for a period of time points to study the progression of some event(s) of interest, some related variables (either time-varying or time-invariant) are also measured over time from the subjects. By jointly modeling the longitudinal outcomes and the time of occurrence of the event(s) of interest, one can (i) study the progression of the outcomes over time, (ii) assess the effects of the longitudinal outcomes on the event-time and (iii) assess the effects of the covariates on the evolution of the longitudinal outcomes and the event-time. In this thesis, we develop different Bayesian models and the computational algorithms for jointly analysing three longitudinal biomarkers and one event-time. Our work is motivated by a clinical experiment conducted by Tata Translational Cancer Research Center, Kolkata, where a group of 236 children, detected as leukemia patients, were treated with two standard drugs (6MP and MTx) nearly for the first two years, and then were followed for the next three years to see if there is a relapse. In our first work we develop a Bayesian joint model for simultaneously imputing the missing biomarker values and for dynamically predicting the non-relapse probability for each patient. In the second work, we develop a Bayesian quantile joint model to assess the effects of the biomarkers on the relapse-time at different quantile levels of the longitudinal biomarkers. Finally, in the third work, we develop a Bayesian latent class joint model for identifying the latent classes with respect to one of the biomarkers and to study the evolution of different biomarkers across different latent clusters. We also dynamically predict the median non-relapse probabilities for different latent classes based on the estimated model parameters. All our works are supported by extensive simulation studies and real applications to leukemia maintenance study.

Acknowledgements

First and foremost I would like to thank my Ph.D. supervisor Dr. Kiranmoy Das for his unconditional support and encouragement during my Ph.D. journey. He taught me how to conduct research and to push my boundaries as I do so. I would forever be grateful to him for never giving up on me, even at low points during my research. Thanks to him for making my Ph.D. journey worthwhile and fulfilling.

I would like to express my heartfelt gratitude to the Director of ISI, the previous and current Dean of Studies, and all previous and current members of the Ph.D. & DSc. Committee (Statistics) and Research Fellow Advisory Committee (Statistics) for believing in me, providing me with the research fellowship, and generously issuing all educational and administrative luxuries, ensuring a smooth and productive Ph.D. tenure at ISI. I would like to thank the previous and present department heads of the Interdisciplinary Statistical Research Unit and Applied Statistical Unit, as well as other distinguished faculty for their help and thoughtful suggestions throughout the years of my Ph.D. I am grateful for the assistance I have gotten throughout the years from the staff members of the library, the Dean's office, the Computer & Statistical Service Centre, and my department.

I would especially want to thank Partha Sarkar, Priya Kedia, Sweata Sen, and Shubhajit Sen, who were former students of my supervisor and with whom I have collaborated on numerous projects. I consider myself lucky to have friends inside and outside of my department who, by their consideration, encouragement, and wisdom, have made my doctoral journey much more pleasurable and less stressful than it would have been without them.

This acknowledgment would not be complete without mentioning the pain staking efforts and patience of my parents. They have always inspired me to pursue greater academic and research endeavours. I owe this thesis to my parents, who have always stood by me and motivated me to pursue this work.

Finally, I want to take this opportunity to thank those working at Tata Medical Centre for their cooperation and generosity in providing us with the rich and diverse dataset that forms the basis of my research.

Damitri Kundu
June, 2023

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AFT	Accelarated Faliure Time
AIC	Akaike Information Criteria
ALD	Asymmetric Laplace Distribution
ALL	Acute Lymphocytic Leukemia
ANC	Absolute Neutrophil Count
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
CI	Credible Interval
CNS	Central Nervous System
DIC	Deviance Information Criterion
HR	High Risk
IG	Inverse Gamma
IR	Intermediate Risk
IW	Inverse Wishart
JM	Joint Model
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
MCMC	Markov Chain Monte Carlo
MCSE	Monte Carlo Standard Error
MH	Metropolis-Hastings
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
MRD	Minimal Residual Disease
NCI	National Cancer Index
PE	Prediction Error

PH	Proportional Hazards
PLT	Platelet count
QR	Quantile Regression
QRJM	Quantile Regression Joint Model
SE	Standard Error
SR	Standard Risk
WBC	White Blood Cell

I dedicate this thesis to my parents

Mrs. Purabi Kundu and Mr. Debabrata Kundu

Chapter 1

Introduction

1.1 Longitudinal Data

Longitudinal data (also known as panel data) are repeated observations of the same variable(s) over different time points. In a balanced data all subjects are measured at all time points with no missing observation. On the other hand, in an unbalanced data (or, irregular setting) different individuals are measured at different time points, and therefore, the number of measurements differ from one individual to the other. Table 1.1 shows the data structure for irregular longitudinal response (Y), that is measured over m time points for n subjects, with "$\times$" denoting a missing observation. Note that in the "Time" column the number of distinct time points is m, but for "Subject 1" and "Subject 4" the total number of measurements are τ_1 and τ_4, respectively, which are smaller than m. In a balanced setting all the columns for subjects would look like the column for "Subject 2" with no missing value. Thus, a general representation of longitudinal response Y, for the i-th subject is given by $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i\tau_i})^T$ measured at time points $(t_{i1}, t_{i2}, \ldots, t_{i\tau_i})^T$, respectively, for $i = 1, \ldots, n$. Longitudinal data are widely observed in various disciplines, including medical studies, biological experiments, agricultural studies, environmental studies, marketing, finance and many others. In medical studies CD4 counts are measured longitudinally for HIV infected people (Wang and Taylor, 2001 [96]); in econometric studies variables measuring the financial health and the physical health of the older individuals are measured longitudinally (Biswas and Das, 2021 [9]); in agricultural studies yield of maize is measured longitudinally across different plots (Wang et al., 2019 [95]); in market studies, ownership of financial products across households are measured over time (Bassi, 2017 [7]); in environmental studies long-term effects of air pollution levels on blood pressure are measured over time (Adar et al., 2018 [2]); in biological experiments tumour growth is measured longitudinally for understanding the dynamics of cancer in mice (Zavrakidis et al., 2020 [104]).

In longitudinal data it is quite common that the observations from the same subject are correlated across different time points, even though the subjects themselves are independent. Thus it makes the traditional approach of modeling each outcome with independent and identically distributed normal invalid. To alleviate this issue,

TABLE 1.1: Longitudinal data structure for the irregular setting with n subjects measured at m time points for a single response variable (Y).

Index	Time	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	$\cdots$	Subject n
1	t_1	Y_{11}	Y_{21}	$\times$	$\times$	Y_{51}	$\cdots$	Y_{n1}
2	t_2	$\times$	Y_{22}	$\times$	$\times$	Y_{52}	$\cdots$	Y_{n2}
3	t_3	Y_{12}	Y_{23}	$\times$	$\times$	Y_{53}	$\cdots$	Y_{n3}
4	t_4	Y_{13}	Y_{24}	Y_{31}	$\times$	Y_{54}	$\cdots$	Y_{n4}
5	t_5	$\times$	Y_{25}	Y_{32}	Y_{41}	Y_{55}	$\cdots$	Y_{n5}
6	t_6	$\times$	Y_{26}	Y_{33}	Y_{42}	Y_{56}	$\cdots$	Y_{n6}
7	t_7	Y_{14}	Y_{27}	Y_{34}	Y_{43}	$\times$	$\cdots$	Y_{n7}
8	t_8	$\times$	Y_{28}	Y_{35}	Y_{44}	$\times$	$\cdots$	Y_{n8}
9	t_9	Y_{15}	Y_{29}	Y_{36}	Y_{45}	$\times$	$\cdots$	Y_{n9}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	t_m	$Y_{1\tau_1}$	Y_{2m}	$\times$	$Y_{4\tau_4}$	$\times$	$\cdots$	Y_{nm}

one could assume $\boldsymbol{Y}_i$ to be independent and identically distributed as $N_{\tau_i}(\boldsymbol{\mu}_i, \Sigma)$, where $\boldsymbol{\mu}_i$ is the $\tau_i \times 1$ mean vector and Σ is covariance matrix of dimension $\tau_i \times \tau_i$. Modeling the mean vector $\boldsymbol{\mu}_i$ is a classical regression problem, where as complexity arises when attempting to model the covariance matrix Σ for which the structure is unknown. In that case the total number of parameters to be estimated is $\frac{\tau_i(\tau_i+1)}{2}$. Even in the balanced cased (i.e., $\tau_i = m$), it is not uncommon to come across the situation where $n < \frac{m(m+1)}{2}$. This leads to "smaller sample larger parameter" in high-dimensional Statistics literature. Laird and Ware (1982) [54] used random effects model and developed a likelihood based estimation approach for modeling longitudinal data. While this approach is handy but it could not explain the unknown covariance structure of the variable(s) of interest. A more flexible approach was proposed in Pourahmadi (1999, 2000) [66] [67], where the author used Cholesky decomposition to ensure the positive definiteness of the estimated covariance matrix. A non-parametric approach of estimating the large covariance matrices was shown in Wu and Pourahmadi (2003) [99]. In Daniels and Pourahmadi (2002) [18], the authors proposed prior distributions that shrink the underlying unknown covariance matrices to some known structures. Pourahmadi's approach was generalized in Pan and Mackenzie (2003) [62] for univariate longitudinal data in irregular setting. Das et al. (2013) [24] used Pan and Mackenzie's (2003) [62] methods for modeling longitudinal biomarkers in gene mapping problem.

For bivariate and/or multivariate setting, the problem is more severe since one has to handle the longitudinal dependence as well as the inter-biomarker dependence over time. A treatment for this issue was proposed in Sy, Taylor and Cumberland (1997) [87] where they used a parametric stochastic model for CD4 T-cells and beta-2

microglobulin in AIDS data. This approach can also handle unbalanced longitudinal data. Theibaut et al. (2002) [88] proposed a linear mixed model for analysing bivariate longitudinal data.

A bivariate autoregressive process was used for detecting prescribing change in two drugs with correlated errors in Sithole and Jones (2007) [86] to model bivariate longitudinal data in regular setting. Das et al. (2011) [22] generalized this approach for the irregular sparse longitudinal data. Random effects models to capture longitudinal dependence for multivariate longitudinal data were proposed in Bandyopadhyay et al. (2010) [6], Ghosh and Hanson (2010) [36] and the references therein.

In longitudinal studies it is not uncommon that the subjects enter the study at different times and/or drop out of it prematurely. This yields unequal number of measurements per subject due to missing values. Following Rubin (1976) [82] the missingness of data is classified into three categories, i.e. (i) missing completely at random (MCAR) (Ibrahim and Molenberghs, 2009 [44]) (ii) missing at random (MAR) (Gabrio et al., 2021 [31]) and (iii) missing not at random (MNAR) (Wang and Hall, 2010 [94]). For the first two cases the dropout is considered as non-informative whereas for the third case it is informative. For handling the informative missing values one has to use the methods proposed in Daniels and Hogan (2008) [17].

1.2 Event-time Data

Event-time data refer to variables that report whether an event of interest has occurred or not, and in addition the time of the occurrence of the event (if any) for each individual.

In event-time data, the event of interest might be observed for a subset of subjects and for the others we will have no data on the event-times since the study ends at some time point. The latter group of individuals are said to be censored at the time point when the study ends. In event-time data we may come across the following situations, (i) we have some dropouts for which we do not have the chance to observe the event of interest (ii) event of interest did not occur for some subjects during the period of study (iii) event has already occurred before the study for some individuals but the exact time of occurrence is unknown (iv) event occurs in certain time interval during the study but the exact time of the occurrence of event is not reported. The first two cases are called right censoring (Lagakos, 1979 [53]), the third one is left censoring (Gomez et al., 1992 [37]) and the last one is called interval censoring (Rodrigues et al., 2018 [81]).

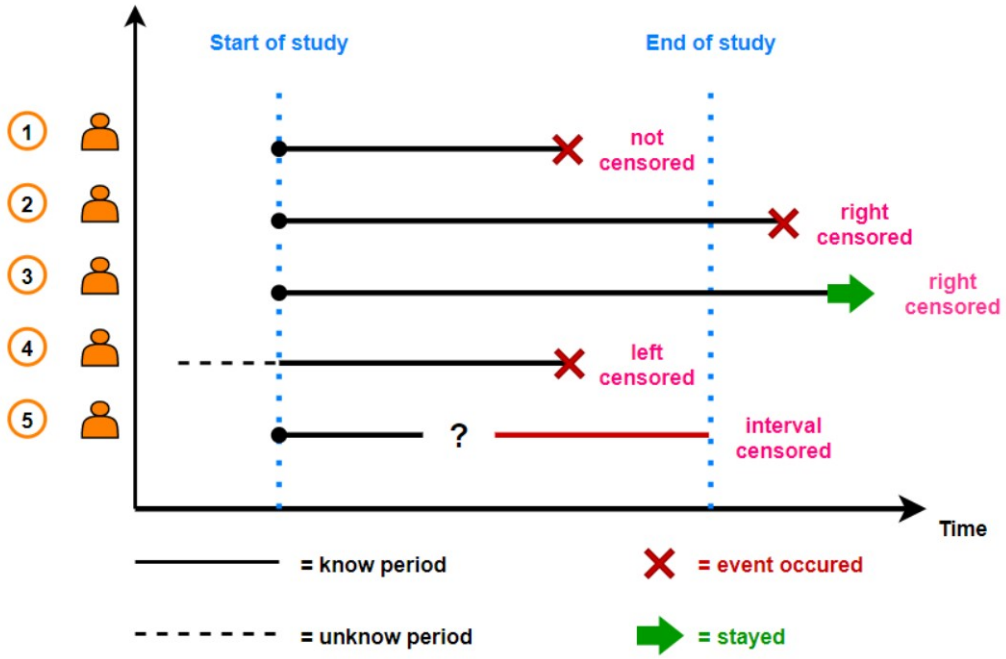


FIGURE 1.1: Different types of censoring for subjects participating in a longitudinal study (here, "stayed" means stayed event-less)

Let T_i and C_i, respectively, denote the actual event-time and the censoring time of the i-th subject. If $T_i < C_i$ then we define the indicator δ_i to be 1 (event-time is observed), otherwise δ_i to be 0 (individual is censored). We define $s_i = T_i \wedge C_i$ and therefore event-time data typically consist of (δ_i, s_i).

Let the probability density function of the true event-time T_i be denoted by $f_{T_i}(.)$. Then $S_i(t)$ is the probability that the event of interest has not occurred until time point t, and we will refer to this quantity as the survival probability as time t and is given by equation (1.1).

$$S_i(t) = Pr(T_i \geq t) = \int_t^\infty f_{T_i}(x)dx;\ t > 0. \tag{1.1}$$

The hazard function, denoted by $\lambda_i(t)$, is defined as follows,

$$\lambda_i(t) = \lim_{\Delta t \to \infty} \frac{S_i(t) - S_i(t + \Delta t)}{\Delta t\ S_i(t)};\ \ t > 0. \tag{1.2}$$

This implies that $S_i(t) = \exp\big(-\int_0^t \lambda_i(x)dx\big).\ \implies\ f_{T_i}(t) = S_i(t)\lambda_i(t)$.

In Statistics literature there are several parametric and non-parametric approaches to model hazard functions. The most commonly used approaches are (i) Cox Proportional Hazards (PH) model (Cox, 1972 [16]) and (ii) Accelarated Failure Time (AFT) model (Zeng and Lin, 2007 [106]); (Mustefa and Chen, 2021 [60]). In this thesis

we consider the PH model since its fits well in our setting of jointly modeling the longitudinal and event-time outcomes. In the following chapters we will perform our analyses based on Cox PH model. A Cox PH model is given as follows:

$$\lambda_i(t) = \lambda_0(t) \times \exp\left(\sum_{p=1}^{P} \beta_p X_{ip}\right), \tag{1.3}$$

where $\lambda_0(t)$ denotes the baseline-hazard, usually modeled by Weibull hazard function (Sahu et al., 1997 [83]), B-spline (Devarajan and Ebrahimi, 2011 [25]; Rizopoulos, 2012 [77]) or wavelets (Moundele et al., 2019 [59]). Here, X_{ip}s are p baseline covariates and βs are the corresponding regression coeffiecients. Both the covariates and the coefficients can either be time-invariant (Verweij and Houwelingen, 1995 [93]) or time-varying (Zhang et al., 2018 [108]; Tian et al., 2005 [89]), as a generalization of the Cox PH model.

1.3 Joint Modeling of longitudinal and event-time data

This thesis will focus on a particular class of models, known as the joint modeling of longitudinal and event-time data. During the past two decades a vast literature has been developed on this class of models mainly due to its practical usefulness and successful applications in various disciplines(mainly in the medical studies). When a group of individuals are followed over time for monitoring the progression of one (or more) event(s) of interest, joint modeling is extremely useful for a powerful Statistical inference. The progression of event(s) typically depends on some outcomes measured longitudinally from the individuals, and therefore, it is of interest to measure the effects of these outcomes on the event-time. In a classical framework, Prentice (1982) used the longitudinal outcomes as time-varying covariates and used a Cox PH model for modeling the event-time. However, since the longitudinal outcomes are measured with some measurement errors this modeling approach results in the biased estimates and hence provides inefficient inference. Moreover, for such setting it is also important to model the trajectories of the biomarkers, and also to assess the effects of the covariates on the progression of the longitudinal outcomes as well as on the event-time. In a joint analysis of longitudinal and event-time data there are three major research interests, i.e. (i) to model the evolution of the longitudinal outcomes, (ii) to assess the effects of the longitudinal outcomes on the hazard function (measuring the instantaneous risk of occurrence of the event), and (iii) to assess the effects of the covariates on the evolution of the longitudinal and the event-time processes. The effectiveness of joint modeling has been established and verified in various papers published in the last two decades.

Joint modeling of a single longitudinal outcome and the time of occurrence of a single event has been proposed in Henderson (2000) [40], Wang and Taylor (2001)

[96], Guo and Carlin (2004) [39]. Models for jointly analysing multiple longitudinal outcomes and a single event-time have been proposed in Lin et al. (2002) [55], Brown et al. (2005) [13], Chi and Ibrahim (2006) [15], Rizopoulos and Ghosh (2011) [79]. These models have also been extended to the occurrence of multiple events (competing events) in Williamson et al. (2008) [97], Hu et al. (2009) [42], Huang et al. (2011) [43]. Also based on such joint models personalized predictive models have been proposed in Proust-Lima and Taylor (2009) [70], Rizopoulos (2011) [76], Tomer et al. (2019) [90], Papageorgiou et al. (2018) [63]. We note that although most of the joint models proposed in the literature are motivated by some medical applications, there are some works where such models are used for gene mapping (Das et al., 2012 [23]) or for some other biological research interests (Das, 2016 [19]). For extensive reviews one may see the review articles by Hogan and Laird (1997) [41], Tsiastis and Davidian (2004) [91], Gould et al. (2015) [38].

Consider a single outcome measured longitudinally at T different time points from a set of N subjects, and let Y_{it} denote the outcome from the i-th subject at the t-th time point. In addition, there are p covariates $x_1, x_2, \ldots, x_p$, which are either time-dependent or time-invariant. For modeling the longitudinal outcome traditionally a linear mixed model is used as follows:

$$Y_{it} = X_{it}^T \boldsymbol{\beta} + Z_{it}^T \gamma_i + \epsilon_{it}, \tag{1.4}$$

where $X_{it} = (x_{1it}, x_{2it}, \ldots, x_{pit})^T$ and $\boldsymbol{\beta}$ is the vector of regression coefficients. Also Z_{it} is the vector (similar to X_{it}) corresponding to the set of covariates with random effects (this set is typically a subset of the set of all predictors) and γ_i is the vector of random effects which capture the longitudinal dependence among the measurements from the same subject collected at different time points. We assume that $\gamma_i \sim N(0, \Sigma)$, and the random errors ϵ_{it}s are iid $N(0, \sigma^2)$.

The hazard function is modeled using the Cox PH model as follows:

$$\lambda_i(t) = \lambda_0(t) \times \exp\left(\psi \mu_{it} + \sum_{p=1}^{P} \delta_p x_{ip}\right), \tag{1.5}$$

where $\lambda_0(t)$ is the baseline hazard function, and $\mu_{it} = X_{it}^T \boldsymbol{\beta} + Z_{it}^T \gamma_i$. The coefficient ψ measures the association between the longitudinal outcome and the hazard. In the medical applications, equation (1.4) models the progression of a biomarker (CD4 count for HIV patients) and the time to an event (death or relapse) is modeled by equation (1.5) which considers the dynamics of the biomarker appropriately. The joint probability distribution of the longitudinal outcomes $\boldsymbol{Y}_i = (Y_{i1}, .., Y_{iT})^T$ and the event-time s_i is then given as follows:

$$f(\boldsymbol{Y}_i, s_i) = \left(\prod_{t=1}^{T} P(Y_{it}|\boldsymbol{\beta}, \gamma_i)\right) \times \left[\{\lambda_i(s_i)\}^{\delta_i} \exp\left(-\int_0^{s_i} \lambda_i(t)dt\right)\right], \tag{1.6}$$

where P denotes the probability distribution of the longitudinal outcome as specified by the model given in equation (1.4). In Chapters 2, 3 and 4, we will introduce different versions of this joint model depending on the problems on interest.

In this thesis, we develop Bayesian models for the joint analysis of multivariate longitudinal outcomes and an event-time with an application to a clinical experiment conducted for the children (from the eastern part of India) detected as leukemia patients.

1.4 Acute Lymphocytic Leukemia

Acute Lymphocytic Leukemia (ALL), also known as Acute Lymphoblastic Leukemia, is quite rare for adults, but it is possibly the most common type of cancer diagnosed among children. It is usually developed from immature white blood cells (WBC) that are the key components to our immune system. ALL is "acute" in the sense that it progresses rapidly by creating immature blood cells. Although ALL is not inherited, it is known that it occurs due to mutations in the DNA (Yokota and Kanakura, 2016 [102]). The most effective drugs used for the treatment of ALL were discovered in the 1960s when the first multi-drug chemotherapy regimens increased its survival rate quite significantly.

Globally, ALL is the main cause of death from cancer among children (Belson et al., 2007 [8]). In 2015, a total of 876,000 confirmed cases of ALL resulting in 110,000 death were globally reported. Even for the developed countries (e.g. the United States) the prevalence of ALL is quite alarming. In 2019, there were (an estimated) 107,620 people living with ALL in the United States as reported by the National Cancer Institute (NCI). Pui and Evans (2013) [71] reported that the survival rate for ALL has increased from 0.10 (in 1960) to 0.90 (in 2010) in most of the developed countries, in particular, in the United States. This is due to success in the maintenance therapy used for treating leukemia patients. Patients are treated for the first one or two years, and are then followed for a longer period (approximately three years) with an expectation that a relapse should not occur after the follow-up period. Pui et al. (2018) [72] reported that the survival rates for ALL are still less than 0.60 in most of the African countries, and are less than 0.70 in most of the Asian countries. More recently, Abdelmabood et al. (2020) [1] analyzed a dataset from Egypt, and estimated the survival rate for ALL as 0.63. They recommend an urgent need for the modification of the current chemotherapy regimens so that the treatments are suitable for local conditions. In fact, they also recommend a better government healthcare globally, and in particular, for Egypt.

Despite the fact that the survival rate for ALL is unsatisfactorily low in India, there is a lack of consensus on its estimate. Approximately, 75,000 new cases of ALL are diagnosed (among children) every year in India, and more than 80% of these

cases occur in families with low income. Arora and Arora (2016) [4] mentioned that a treatment abandonment (which typically occurs due to the lack of financial support) may result in the lack of data for which the survival rate cannot be estimated accurately in India. Varghese et al. (2018) [92] analyzed data from the Southern part of India, and reported the overall survival rate for ALL as 0.40; with an event-free survival rate as 0.28. Note that "event-free survival" refers to survival without a relapse (recurrence of cancer) or other related complications.

Most of the existing works on joint modeling (with medical applications) focus on the survival (no death) of the patients, which is of course an important event to note. However, for assessing the effectiveness of the treatment it is equally important to note the time-to-relapse (if any). Multiple recurrences definitely result in adverse effects on the patients' health conditions (and result in death eventually), and therefore it is extremely important to identify the biomarkers associated with the relapse time, and also to assess the effects of the drugs on those biomarkers.

1.4.1 Motivating Dataset

Tata Translational Cancer Research Center (TTCRC) in Kolkata conducted a study in which 236 children, suffering from ALL, were treated with advanced chemotherapy. The study started in 2014 and ended in 2019. As a part of the Tata Medical Center (TMC), TTCRC develops advanced treatments for the cancer patients in the eastern part of India. In the current study, children (patients) were treated with two standard drugs, i.e. 6-mercaptopurine (6MP) and methotrexate (MTx). The median duration of the treatment was 92 weeks. Ethical approval was obtained from TMC-Institutional Review Board.

The starting time and the end time of the study were fixed, but all the 236 children did not join the program at the same time point, and the number of visits were also different for different children. Based on the total WBC count (at presentation) and the risk-group specified by the National Cancer Institute (NCI), patients were treated with different drug doses at the beginning. After that the subjects were tested for WBC count, ANC, and platelet count in the subsequent time points (bi-weekly, or once in a month). The median number of visits was 41. After the end of the treatment patients were followed at most for the next 178 weeks. If there was a relapse either during treatment or in the follow-up period then the time-to-relapse was noted; otherwise the subject was censored at the time when the follow-up ended for that patient.

Among 236 patients, 36% were female and 64% were male; and the children belonged to the age interval [1-17.5] with a median age of 4.7 years (at presentation). Only 29% of the children had a bulky disease (i.e. the cancerous masses 10 cm or larger in diameter), and 31% had a disease related to the central nervous system

(CNS). Almost 71% of the patients did not show the sign of a relapse when the study ended in May, 2019.

For our modeling, drug doses are the time-varying covariates with fixed effects. Among the time-invariant covariates, we have (i) age at diagnosis, (ii) gender (M/F), (iii) lineage (B cell/T cell), (iv) WBC count at presentation, (v) NCI risk group (high risk/standard risk), (vi) presence of bulky disease (Y/N), (vii) presence of CNS disease (Y/N), (viii) risk at presentation, (ix) day 8 risk, (x) day 35 risk, (xi) morphological remission (Y/N/Unknown), and (xii) minimal residual disease (MRD) status. We note that NCI "high-risk" group refers to the children with WBC counts higher than 50,000 (cells/mm^3), and "standard-risk" group refers to the group with counts lower than 50,000 (cells/mm^3), at presentation. By the end of the treatment phase, if ANC and platelet count are in the normal range for a patient then the patient is considered in morphological complete remission (CR). The MRD refers to the remaining cancer cells after treatment. Table 1.2 provides the list of variables used in our modeling, and their roles in our analysis. The summary statistics of the time-invariant covariates are provided in Table 1.3.

In our dataset, there are some missing biomarker values for some patients. Overall percentage of missingness for WBC count, neutrophil count and platelet count are 32.04, 4.83 and 4.80, respectively. We note that there is no withdrawal in the study, and hence these missing values are not because of the dropouts. In fact, we observe missingness only in the longitudinal biomarkers and not in the covariates. As informed by TTCRC, these biomarker values are missing for no valid reason and mostly because of the human error. Therefore, we need to improve our estimation method without really worrying about why these responses are missing.

In Figure 1.2 we plot the longitudinal trajectories of the three biomarkers for four randomly selected patients. We notice that for the patients with a relapse (in the follow-up period) the biomarkers are mostly lower than the respective grand means (shown by the solid lines) during the treatment. For the patients with no relapse, the biomarkers are mostly above the respective grand means. This indicates that a relapse is possibly associated with the observed values of the biomarkers.

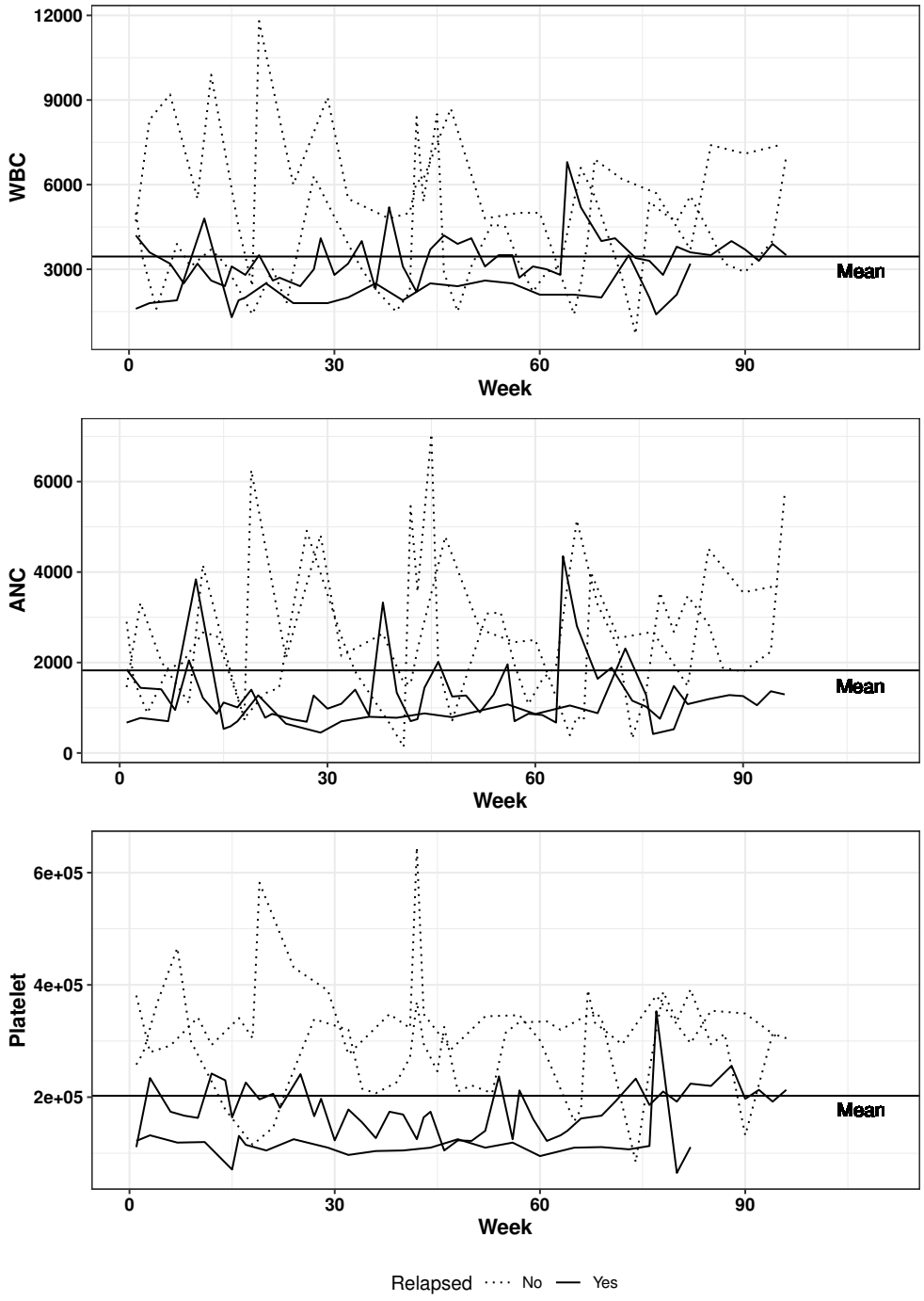


FIGURE 1.2: Longitudinal trajectories of the three biomarkers for four randomly selected children. Solid lines represent trajectories for the children with a relapse (in the follow-up period), and the dotted lines are for those with no relapse in the follow-up period.

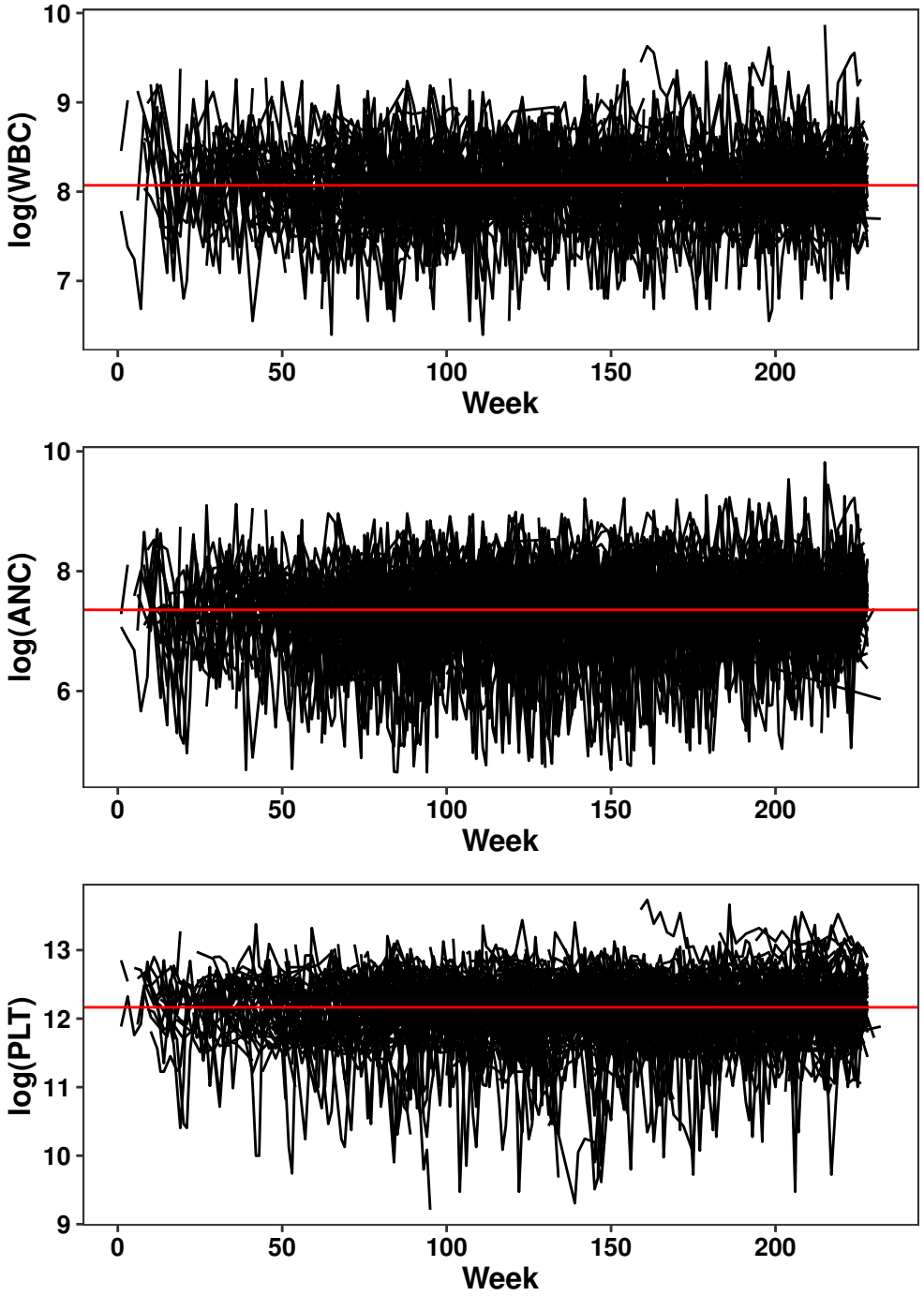


FIGURE 1.3: Log tansformed longitudinal trajectories of WBC, neutrophil (ANC) and platelet (PLT) count of the patients

TABLE 1.2: Variables used in the ALL data analysis. The role of each variable in our model is also specified.

Name	Type	Role (in our model)
White blood cell count	Continuous	Outcome
Neutrophil count	Continuous	Outcome
Platelet count	Continuous	Outcome
Time to Relapse	event-time	Outcome
Dose of 6MP	Continuous	Time-varying covariate
Dose of MTx	Continuous	Time-varying covariate
Age at diagnosis	Continuous	Fixed Covariate
WBC count at presentation	Continuous	Fixed Covariate
Gender	Binary	Fixed Covariate
Lineage	Categorical	Fixed Covariate
NCI risk group	Categorical	Fixed Covariate
Bulky disease	Binary	Fixed Covariate
CNS disease	Binary	Fixed Covariate
Risk at presentation	Categorical	Fixed Covariate
Day 8 risk	Categorical	Fixed Covariate
Day 35 risk	Categorical	Fixed Covariate
Morphological Remission	Categorical	Fixed Covariate
MRD status	Categorical	Fixed Covariate

TABLE 1.3: Summary statistics for the time-invariant covariates in the ALL dataset.

Variable	Summary
Age at diagnosis	Min= 1, Q1=3.091, Median=4.7, Q3=8.292, Max= 17.5
WBC count at presentation	Min=100, Q1=7175, Median=15910, Q3= 42300, Max= 983500
Gender	Female: 36%, Male: 64%
Lineage	B cell: 85%, T cell: 15%
NCI risk group	High Risk: 36%, Standard Risk: 64%
Bulky disease	Yes: 29%, No: 69%, Unknown: 2%
CNS disease	Yes: 31%, No: 64%, Unknown: 5%
Risk at presentation	High Risk: 24%, Standard Risk: 46%, Intermediate Risk: 30%
Day 8 risk	High Risk: 30%, Standard Risk: 38%, Intermediate Risk: 30%, Other: 2%
Day 35 risk	High Risk: 43%, Standard Risk: 27%, Intermediate Risk: 23%, Other: 7%
Morphological Remission	Yes: 94%, No: 3%, Unknown: 3%
MRD status	Positive: 16%, Negative: 67%, T cell: 7%, Other: 10%

1.4.2 Data Processing

In our analysis we have considered those subjects, with no missing value in the fixed covariates and medicine doses. We have also removed data for the subjects with less than 5 non-missing observations for longitudinal outcomes. We consider the log transformed longitudinal biomarkers for stabilizing the variances and then were centered with respect to the respective median values.

Variables corresponding to time (week, observed relapse-time and censoring time) were brought to the same scale, and the continuous covariates such as the age at diagnosis and log transformed WBC count at presentation were normalized with respect to their observed means and standard deviations. Figure 1.3 shows that the longitudinal responses are variance-stabilized after being log transformed.

1.5 Major Contributions and Inferential Objectives

The dataset under consideration is quite challenging in the sense that there are significant percentage of missing outcomes. In a joint modeling imputation of the missing outcomes is computationally demanding since the models are already quite complex. We implement a simpler imputation technique that automatically imputes the missing outcomes within each iteration of MCMC algorithm by appropriately considering the dependence within and between the outcomes. In addition, since the dataset is not very large (which is usually the case in biomedical researches) we use posterior predictive distributions for better inference. All these approaches are presented in **Chapter 2**.

While there is a wealth of literature on joint modeling of longitudinal outcomes and time-to-event there are limited works on quantile-based modeling, mainly due to the computational complexities involved in such models. We develop a multivariate quantile regression model for jointly modeling multiple longitudinal outcomes and time-to-event. Since quantile-based inference is more accurate and provides a complete picture of the effects of covariates on the outcomes, our model and analysis reported in **Chapter 3** is a real contribution to the existing literature.

In **Chapter 4**, we develop a latent class Bayesian joint model. While such models are routinely used for analyzing multivariate longitudinal data the interpretation of the latent classes are less obvious for most of the existing works. We develop a novel approach of clustering the most important outcome (in this case, the lymphocyte count) and then assessing the other outcomes with respect to these clusters. This approach, although models multivariate outcomes, selects only one of those for clustering, and thus the interpretation of the latent clusters are very clear. In addition, such approach is computationally faster than the other clustering algorithms popularly used in multivariate analysis.

1.6 Organization of the thesis

The work presented in this thesis is organized in different chapters, and is summarized as follows:

In **Chapter 2**, we develop a Bayesian joint model in which a linear mixed model is used to jointly model three biomarkers (i.e. white blood cell count, neutrophil count, and platelet count) and a semi-parametric proportional hazards model is used to model the relapse-time. Our proposed joint model can assess the effects of different covariates on the progression of the biomarkers, and the effects of the biomarkers (and the covariates) on relapse-time. In addition, the proposed joint model can impute the missing longitudinal biomarkers efficiently. Our model can also dynamically predict the non-relapse probabilities for each patient based on the historical data. Our analysis shows that the white blood cell (WBC) count is not associated with relapse-time, but the neutrophil count and the platelet count are significantly associated with it. We also infer that a lower dose of 6MP and a higher dose of MTx jointly result in a lower relapse probability in the follow-up period. Interestingly, we find that relapse probability is the lowest for the patients classified into the "high-risk" group at presentation. The effectiveness of the proposed joint model is assessed through the extensive simulation studies.

Chapter 3 presents the importance of using quantiles over the mean in jointly modeling the longitudinal biomarkers and the event-time. Linear mixed models are traditionally used for jointly modeling longitudinal outcomes and event-time(s). However, in the presence of some time-varying covariates it might be of interest to see how the effects of different covariates vary from one quantile level (of outcomes) to the other, and consequently how the event-time changes across different quantiles. For such analyses linear quantile mixed models can be used into the joint modeling framework, and an efficient computational algoithm can be developed. Quantile based analysis is also appropriate when the joint distribution of the biomarkers deviates from a multivariate normal distribution. We consider an Asymmetric Laplace Distribution (ALD) for each outcome, and exploit the mixture representation of the ALD for developing an efficient Gibbs sampler algorithm for the proposed linear quantile joint regression model. A multivariate Brownian motion is considered for the subject-specific random effects for higher flexibility. From our analysis we infer that a higher lymphocyte count accelerates the chance of a relapse while a higher neutrophil count and a higher platelet count (jointly) reduce it. Also, we infer that across (almost) all quantiles 6MP reduces the lymphocyte count, while MTx increases the neutrophil count. Simulation studies are performed to assess the effectiveness of the proposed approach.

In **Chapter 4**, we present a Bayesian latent class joint model. In a joint modeling framework we use a latent class model for the lymphocyte count since it is the most

important biomarker associated to ALL. The other two biomarkers, i.e. the neutrophil count and the platelet count are modeled using linear mixed models, and the event-time is modeled by the semi-parametric proportional hazards model. The model parameters are estimated by the Markov Chain Monte Carlo (MCMC) algorithm. Our analysis detects two latent classes for the lymphocyte count, and we estimate the Kaplan-Meier functions of the non-relapse for both these groups. Our simulation studies illustrate the discriminating and the predictive power of the proposed approach compared to the usual mean regression based traditional joint models.

Finally in **Chapter 5**, we summarize the work presented in this thesis. We highlight the major methodological contributions, and the interesting inferences based on our data analysis. We also discuss some limitations of our work, and also discuss the possibilities of further extending our work from some different perspectives.

Chapter 2

A Bayesian joint model for multivariate longitudinal and event-time data

2.1 Preamble

There is a vast literature on the joint modeling of longitudinal outcomes and event-time in the last two decades mainly due to the meaningful practical applications of such models in biomedical studies. In the existing literature for joint modeling we find several ways to jointly modeling these two types of responses, each with their own advantages and disadvantages. One approach of joint modeling involves in assuming the longitudinal responses to be measured without errors, so that the responses can be used as time-varying covariates in the extended Cox PH model (Andersen and Gill, 1982 [3]) for modeling the event-time process. This method, while simple, comes with its own problems i.e., (i) it is unrealistic for longitudinal measurements to be measured without error. For example, the machine used for monitoring say, the blood pressure of an individual over time, can be faulty, (ii) in many studies the event of interest (relapse) is usually observed much later than the cessation of the longitudinal study, in which case the longitudinal response at or near the event-time points can be at best set as the last observations of the longitudinal response, (iii) it is also quite common to notice dropouts in longitudinal study which are often non-random (Wang and Taylor, 2001 [96]), thus the assumption of longitudinal response being covariate in the Cox PH model prevents the model from acknowledging its association to the underlying failure mechanism (Kalbfleisch and Prentice, 2002 [46]; Prentice, 1982 [68]). All of the above problems lead to an increased bias in the estimated parameters. The last value carried forward approach which can handle the second problem, is quite unrealistic since it is unlikely that the longitudinal response remains unchanged for a long period after a certain time point. This issue was discussed in Tsiatis and Davidian (2004) [91] where the authors suggested the imputation of longitudinal responses by modeling them using a mixed effects model. The authors focused particularly on joint models with shared random effects for the

longitudinal and the event-time process. Another way of joint modeling would be to model the longitudinal responses using the mixed effects model, and then separately model the event-time process with an extended Cox PH model where expected part of longitudinal responses would serve as time-varying covariates. This type of modeling was referred to as separate models in Wang and Taylor (2001) [96]. This solves the first two problems but does not solve the third one. In the context of separate modeling, not accounting for the third problem, increases bias in the estimates of the model parameters (for the longitudinal model), which when substituted later as covariate in the extended Cox PH model increases the bias in the estimates. All of the above reasons serve as motivation for the joint modeling shown in equation (1.4) and (1.5) of Chapter 1.

Henderson et al. (2000) [40] proposed different likelihood based models for such joint modeling. In a Bayesian setting, Wang and Taylor (2001) [96] developed a joint model for the CD4 counts and time to progress into AIDS for HIV patients. Guo and Carlin (2004) [39] considered similar Bayesian models for comparing efficacy of two antiretroviral drugs. Fieuws and Verbeke (2004) [30], Chi and Ibrahim (2006) [15], Rizopoulos and Ghosh (2011) [79], Zhu et al. (2011) [109], developed joint models for multivariate longitudinal and survival data. Rizopoulos (2011) [76], Rizopoulos et al. (2017) [80] developed flexible Bayesian joint models which can automatically predict the subject-wise survival probability over time. All these authors considered (generalized) linear mixed models for the longitudinal outcomes, and a proportional hazards (PH) or an accelerated failure time (AFT) model for the time-to-event, and then link the two models either by shared random effects or by correlated random effects.

We build our work on the existing literature, and develop a Bayesian joint model for jointly analyzing WBC count, absolute neutrophil count (ANC), platelet count; and time-to-relapse. The motivation for developing a joint model for our dataset is discussed in Section 1.4.1. We consider a number of joint stochastic models where different Gaussian correlated random effects are used to capture the longitudinal and the inter-biomarker dependence. By using some popular model selection criteria we choose the "best fit" model for our dataset. The missing (correlated) biomarker values are imputed multiple times within each iteration of the Markov Chain Monte Carlo (MCMC) algorithm using the working joint model, and then probability of a relapse is predicted for each subject for the follow-up period (after the end of treatment). Our analysis addresses some clinically interesting research questions; for example, how the trajectories of different biomarkers will be different (on the average) for the patients receiving a lower dose of 6MP and a higher dose of MTx from the patients receiving a median dose of each drug.

Based on the three types of available data (i.e. relapse-time, longitudinal biomarkers, and the covariates) it is of interest to study (i) the progression of the biomarkers with time, and the effects of the covariates on it, (ii) effects of the biomarkers on

time-to-relapse, and (iii) (direct) effects of the covariates on time-to-relapse. A joint model for the longitudinal biomarkers and time-to-relapse is used here for such inference. We also use the posterior predictive distribution for assessing the effects of some important predictors (for example, gender, risk group, medicine dose etc.) on the longitudinal process as well as on the event-time.

The rest of this chapter is organized as follows. In Section 2.2 of Chapter 1, we describe the proposed Bayesian joint models. The computational details are also discussed in this section. The results from the data analysis are summarized in Section 2.3. In Section 2.4, we report the results from simulation studies. Finally in Section 2.5, we summarise our work.

2.2 Proposed Model and Estimation Method

Recall that we have three biomarkers, i.e. (i) WBC count, (ii) neutrophil count (ANC), and (iii) platelet count, which we consider as longitudinal outcomes. For stabilizing the variances in the raw biomarker values, we consider log-transformed biomarker values for our analysis (refer to Figure 2.1). Let Y_{ijk} be the (log transformed) k-th biomarker ($k = 1, 2, 3$) from the i-th patient at time t_{ij}, for $j = 1, 2, \ldots, \tau_i$. Note that the number of longitudinal measurements differs from one patient to the other, and hence we use the notation τ_i to denote the number of measurements from the i-th patient. For each patient we either observe the relapse time (T_i) when the cancer returns (during treatment or in the follow-up), or the censoring time (C_i) when the follow-up ends for the i-th patient. We define an indicator variable $\delta_i = 1$, if $T_i < C_i$; (and 0, otherwise), and define s_i =min(T_i, C_i) as the time-to-event for the i-th patient.

2.2.1 Longitudinal Submodel

We propose the following multivariate linear mixed model for the longitudinal biomarkers. Our model is similar to the models proposed in Henderson et al. (2000) [40], Rizopoulos and Ghosh (2011) [79], Das (2016) [19]:

$$Y_{ijk} = f_k(t_{ij}) + \boldsymbol{\beta}_{1k}^T \mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^T \mathbf{z}_i + W_{ik}(t_{ij}) + e_{ijk}, \tag{2.1}$$

where the continuous function $f_k(t_{ij})$ is the general effect of time on the k-th biomarker, and we model it as a polynomial function of time, i.e. we consider $f_k(t_{ij}) = \eta_{k0} + \eta_{k1} t_{ij} + \eta_{k2} t_{ij}^2 + \ldots + \eta_{kr} t_{ij}^r$. We note that some unknown functions of time could be used for modeling $f_k(t_{ij})$, but the plots (in Figure 2.1) suggest that a polynomial function should suffice for our dataset. For selecting the optimal order (r) of the polynomial function, we use the deviance information criteria (DIC) for a linear mixed model as proposed in Celeux et al. (2006) [14]. The vectors $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$,

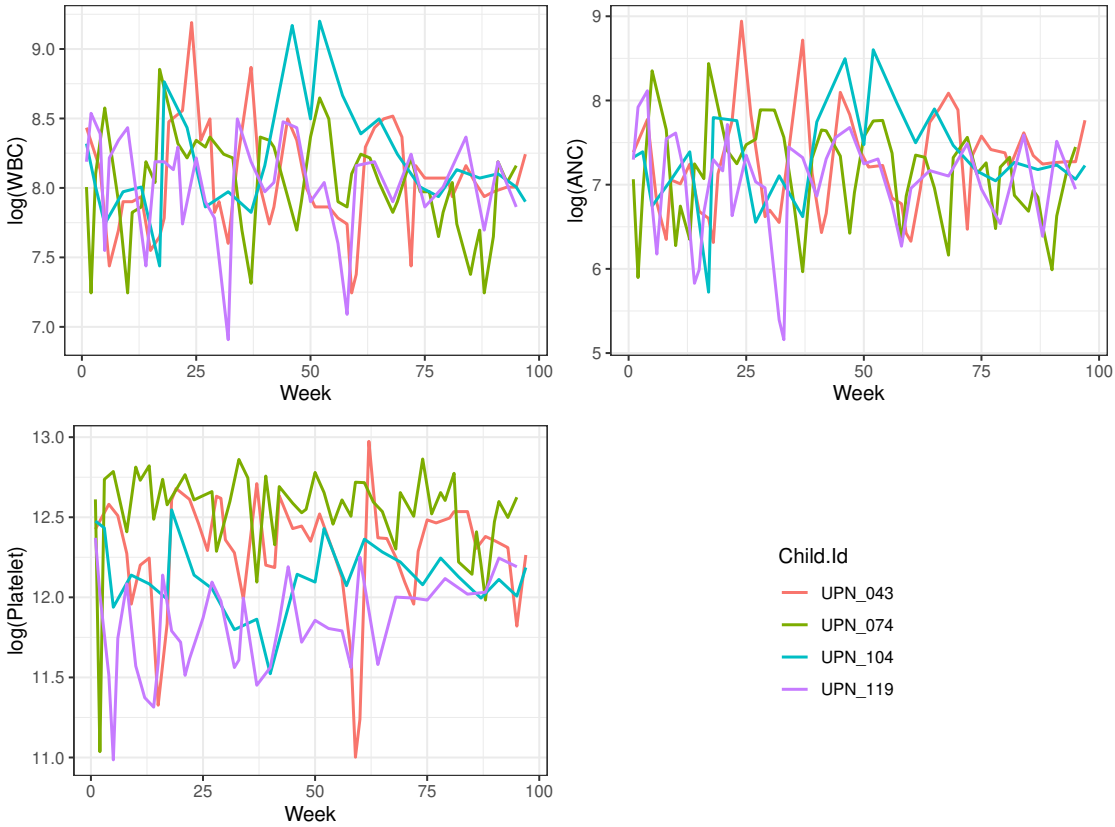


FIGURE 2.1: Longitudinal (log-transformed) biomarkers (WBC count, ANC, and Platelet count) for four randomly selected subjects in ALL dataset.

respectively, denote the (fixed) effects of the time-varying covariates ($\mathbf{x}_{ij}$), and the time-invariant covariates ($\mathbf{z}_i$) on the k-th biomarker. Recall that the doses of two medicines are taken as time-varying covariates, while there are several other time-invariant covariates as discussed in Table 1.2 in Chapter 1. The zero-mean Gaussian random effects $W_{ik}(t_{ij})$ capture the longitudinal dependence as well as the dependence among the three biomarkers (Rizopoulos, 2016 [78]). Note that these random effects are biomarker-specific since the between-patient variations might be different for different biomarkers. The random errors e_{ijk} are assumed to be iid N(0,σ^2_{ek}), and are independent to $W_{ik}(t_{ij})$.

We consider two different specifications for $W_{ik}(t_{ij})$ in our analysis, following the existing literature. First, we consider model with random intercepts only, i.e. $W_{ik}(t_{ij}) = a_{ik}$, where $\mathbf{a}_i = [a_{i1}, a_{i2}, a_{i3}]^T \sim N(0, D_1)$, and D_1 is a 3 $\times$ 3 covariance matrix. This specification assumes that the dependence among the biomarkers and the longitudinal dependence (for each biomarker) remain unchanged throughout the study.

As an alternative specification, we consider model with random intercepts and random slopes (of time), i.e. $W_{ik}(t_{ij}) = a_{ik} + b_{ik}t_{ij}$, where $[\mathbf{a}_i^T, \mathbf{b}_i^T]^T \sim N(0, D_2)$, and

$\mathbf{b}_i = [b_{i1}, b_{i2}, b_{i3}]^T$. Here D_2 is a 6×6 covariance matrix, and this specification assumes that the inter-biomarker dependence and the longitudinal dependence change with time.

2.2.2 Time-to-Event Submodel

Since the time-to-relapse is (possibly) associated with the longitudinal biomarkers, a joint model is meaningful in our analysis (refer to Figure 1.2). However, the nature of this association can be complicated, and therefore many different specifications of the joint model exist in the literature (Henderson et al., 2000 [40]; Rizopoulos and Ghosh, 2011 [79]; Das, 2016 [19]; and the references therein). We consider Cox proportional hazards (PH) model, which is the most commonly used model for time-to-event data. We consider two alternative specifications of PH model for linking it to the longitudinal submodel given in equation (2.1).

In our first choice, we use the expected longitudinal outcomes (conditional on the random effects) as time-varying predictors in the PH model. Note that the model in equation (2.1) can also be written as follows: $Y_{ijk} = \mu_{ik}(t_{ij}) + e_{ik}(t_{ij})$, where $\mu_{ik}(t_{ij}) = f_k(t_{ij}) + \boldsymbol{\beta}_{1k}^T \mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^T \mathbf{z}_i + W_{ik}(t_{ij})$. For a fixed time point, say t, we define $\boldsymbol{\mu}_i(t) = [\mu_{i1}(t), \mu_{i2}(t), \mu_{i3}(t)]^T$. Let $\lambda_i(t)$ denote the hazard (instantaneous probability of relapse) for the i-th patient at time t. Assuming that the expected biomarker values (conditional on the Gaussian random effects) are associated with hazards, we consider the following PH model:

$$\lambda_i(t) = \lambda_0(t) \exp\left[\Psi_1^T \boldsymbol{\mu}_i(t) + \boldsymbol{\theta}_1^T \mathbf{z}_i\right], \tag{2.2}$$

where the vector of coefficients Ψ_1 measures the effects of the expected longitudinal biomarkers on the hazards (Das, 2016 [19]). And $\boldsymbol{\theta}_1$ denotes the effects of the fixed covariates on hazards, λ_0 denotes the baseline hazard. This specification assumes that the time-to-relapse depends on the expected biomarker values, and on the fixed covariates. The effects of the two drugs on the time-to-relapse are rather indirect, i.e. only through the observed biomarkers.

As an alternative specification, we also consider the following PH model for the time-to-relapse:

$$\lambda_i(t) = \lambda_0(t) \exp\left[\Psi_2^T x_i^* + \boldsymbol{\theta}_2^T \mathbf{z}_i + W_i^*(t)\right], \tag{2.3}$$

where x_i^* is a 2×1 vector of the median dose (during the treatment) for 6MP and MTx given to the i-th patient, and the vector Ψ_2 measures the effect of the median doses on time-to-relapse. Note that we consider the median dose as a covariate since the median gives a robust summary of the patient's dose distribution. This specification assumes that the drugs and the time-invariant covariates directly affect the time-to-relapse. Similar to the model in equation (2.2), $\boldsymbol{\theta}_2$ denotes the effects of the fixed covariates on hazards. Here $W_i^*(t)$ are zero-mean Gaussian random effects; different

specifications for $W_i^*(t)$ are considered to select the "best fit" model for our data (discussed in Section 2.3.2). The dependence between the longitudinal biomarkers and the time-to-relapse can be modeled by considering a joint distribution for $W_{ik}(t_{ij})$ and $W_i^*(t)$, as described in Section 2.3.2 (for Models III and IV).

For modeling the base-line hazard function $\lambda_0(t)$ in equations (2.2) and (2.3), we use flexible cubic B-splines following the JMbayes package in R, which we use for our computation. This package was written by Rizopoulos (2016) [78], and has been used for Bayesian joint modeling by several authors (Balan and Putter, 2020 [5]; Papageorgiou et al., 2019 [64]). This package models $\lambda_0(t)$ as follows: $log\lambda_0(t) = \gamma_{\lambda_0,0} + \sum_{q=1}^{Q} \gamma_{\lambda_0,q} B_q(t,v)$, where $B_q(t,v)$ denotes the q-th basis function of B-spline with knots $v_1, v_2, \ldots, v_Q$. For detection of the optimal number (and the location) of knots the JMbayes package considers a relatively large number of knots (15, 20, 30 etc.), and then penalize the B-spline coefficients by considering suitable prior distributions (e.g. Laplace prior, Horseshoe prior). The non-zero coefficients are finally considered in the model.

2.2.3 Joint Likelihood and Estimation Method

For any specification of the submodels, let $\boldsymbol{\alpha}_i$ denote the vector of subject-specific random effects. For example, if we specify $W_{ik}(t_{ij}) = a_{ik} + b_{ik}t_{ij}$, and consider equation (2.2) as the submodel for the time-to-event data, then $\boldsymbol{\alpha}_i = [\mathbf{a}_i^T, \mathbf{b}_i^T]^T$, and $\boldsymbol{\alpha}$ denotes the vector of subject-specific random effects from all subjects. Additionally, let $\boldsymbol{\Theta}$ denote the set of all fixed model parameters, and $\boldsymbol{\beta}$ denotes the set of fixed model parameters in the longitudinal submodel . The complete data likelihood is expressed as follows:

$$L(\boldsymbol{\Theta}|Y,S,\boldsymbol{\alpha}) = \prod_{i=1}^{236}\left(\prod_{j=1}^{\tau_i}\prod_{k=1}^{3} P(Y_{ijk}|\boldsymbol{\beta},\boldsymbol{\alpha}_i)\right) \times \left[\{\lambda_i(s_i|\boldsymbol{\Theta},\boldsymbol{\alpha}_i)\}^{\delta_i} \exp\left(-\int_0^{s_i}\lambda_i(t)dt\right)\right] \times \pi(\boldsymbol{\alpha}_i), \tag{2.4}$$

where P denotes the probability distribution of the biomarker Y_{ijk} conditional on $\boldsymbol{\alpha}_i$ from equation (2.1), and $\pi(\boldsymbol{\alpha}_i)$ denotes the probability distribution of $\boldsymbol{\alpha}_i$. We use the likelihood given in equation (2.4), and by considering appropriate prior distributions for the model parameters we compute the joint posterior distribution. All our inferences are based on the joint posterior distribution. We sample from the joint posterior distribution using a hybrid combination of Gibbs sampler and Metropolis-Hastings algorithm, and estimate the model parameters by their respective sample means. All our computations are done using JMbayes package in R (Rizopoulos, 2016 [78]).

In our dataset, there are some missing biomarker values for some patients. Since the missing observations (only in the biomarkers and not in the covariates) are purely

due to human error (as reported by TTCRC), we assume an "ignorable" missingness. Instead of considering only the complete data (data for the patients with no missing values) or the available data, we impute the missing outcomes within MCMC iterations using the proposed joint model for improving the estimates of the model parameters. This imputation inherently assumes "missing at random" (MAR) as the missingness mechanism, and imputes the unobserved biomarkers as follows.

In the m-th iteration of MCMC, let $\Omega^{(m)}$ denote the set of updated (estimated) model parameters. Conditional on $\Omega^{(m)}$, we first sample the subject-specific random effects $\boldsymbol{\alpha}_i$ for the i-th patient. Then we sample the missing biomarker(s) from the current step's predictive distribution(s) conditional on the observed biomarkers, covariates, and the random effects. Since we have multivariate longitudinal biomarkers, we need to condition on the random effects for sampling correlated data (Schafer and Yucel, 2002 [84]). This is done for all the T time points, and thus we get a complete dataset (with no missing biomarkers). This complete dataset is used for estimating model parameters in the $(m+1)$-th iteration. Thus, in each iteration we update the parameters and the missing biomarkers simultaneously. By considering M iterations, we thus get a set of M complete datasets based on which we get the final estimates (as the average of the estimates from each dataset) of the model parameters.

2.2.4 Subject-wise Prediction of Relapse Probabilities

The proposed joint model is used for dynamic prediction of the subject-wise relapse probability based on the historical longitudinal measurements and covariates (Rizopoulos, 2016 [78]). Let $H_i(t)$ denote historical measurements for the i-th subject who is event-free (no relapse) upto time point t, i.e. $H_i(t) = \{Y_{ijk}, x_{ij}, z_i; 0 \leq j \leq t, k = 1,2,3\}$. Additionally, let D_n denote the training dataset, i.e. $D_n = \{Y_{ijk}, x_{ij}, z_i, s_i, \delta_i;\ k = 1,2,3;\ j = 1,2,\ldots,\tau_i;\ i = 1,..,236\}$. Given that the subject is event-free (no relapse) until time t, the probability that it will be event-free upto time $u = t + \Delta t$, for $\Delta t > 0$, is given as follows:

$$\begin{aligned} \pi_i(u|t) &= Pr\big(T_i \geq u | T_i > t, D_n, H_i(t)\big) \\ &= \int Pr\big(T_i \geq u | T_i > t, H_i(t), \boldsymbol{\Theta}\big) Pr\big(\boldsymbol{\Theta}\big|D_n\big) d\boldsymbol{\Theta}, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\Theta}$ denotes the set of all fixed model parameters in the joint model (equation 2.4). Additionally,

$$Pr\big(T_i \geq u | T_i > t, H_i(t), \boldsymbol{\Theta}\big) = \int \frac{S_i\big[u|\tilde{H}_i(u, \boldsymbol{\alpha}_i, \boldsymbol{\Theta}), \boldsymbol{\Theta}\big]}{S_i\big[t|\tilde{H}_i(t, \boldsymbol{\alpha}_i, \boldsymbol{\Theta}), \boldsymbol{\Theta}\big]} Pr(\boldsymbol{\alpha}_i | T_i > t, H_i(t), \boldsymbol{\Theta})\ d\boldsymbol{\alpha}_i, \tag{2.6}$$

where $\tilde{H}_i(u, \boldsymbol{\alpha}_i, \boldsymbol{\Theta})$ denotes the estimated subject-specific longitudinal trajectories (with random effects) until time point u, and $S_i(.)$ is the event time function conditional on $\tilde{H}_i$. Here, $Pr(\boldsymbol{\alpha}_i | T_i > t, H_i(t), \boldsymbol{\Theta})$ denotes the posterior distribution of the

random effects, and $Pr(\boldsymbol{\alpha}_i|T_i > t, H_i(t), \boldsymbol{\Theta}) \propto P(T_i > t|\boldsymbol{\alpha}_i, \boldsymbol{\Theta})P(H_i(t)|\boldsymbol{\alpha}_i, \boldsymbol{\Theta})P(\boldsymbol{\alpha}_i|\boldsymbol{\Theta})$.

A Monte Carlo estimate of $\pi_i(u|t)$ is obtained as follows. Sample $\tilde{\boldsymbol{\Theta}}^l$, $l = 1, 2, \ldots, L$; from $Pr(\boldsymbol{\Theta}|D_n)$ as post burn-in iterations. For a fixed $\tilde{\boldsymbol{\Theta}}^l$, we sample $\tilde{\boldsymbol{\alpha}_i}^{l_q}$ ($l_q = 1, 2, \ldots, L_q$) from $Pr(\boldsymbol{\alpha}_i|T_i > t, H_i(t), \tilde{\boldsymbol{\Theta}}^l)$, and compute

$$\hat{\pi_i}^l(u|t) = \frac{1}{L_q}\sum_{l_q=1}^{L_q} \frac{S_i\big[u|\tilde{H}_i(u, \tilde{\boldsymbol{\alpha}_i}^{l_q}, \tilde{\boldsymbol{\Theta}}^l), \tilde{\boldsymbol{\Theta}}^l\big]}{S_i\big[t|\tilde{H}_i(t, \tilde{\boldsymbol{\alpha}_i}^{l_q}, \tilde{\boldsymbol{\Theta}}^l), \tilde{\boldsymbol{\Theta}}^l\big]}. \tag{2.7}$$

And finally, we compute

$$\hat{\pi_i}(u|t) = \frac{1}{L}\sum_{l=1}^{L} \hat{\pi_i}^l(u|t), \tag{2.8}$$

based on all L samples, and plot the non-relapse probabilities over time.

2.3 Data Analysis

2.3.1 Prior specifications and computational details

We specify diffuse prior distributions for the model parameters following the existing literature (Das, 2016 [19]; Rizopoulos, 2016 [78]). For $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$ in equation (2.1), we consider multivariate normal prior with mean vector=0, and diagonal covariance matrices with the diagonal element=1000. For the residual variances σ^2_{ek} we specify an Inverse Gamma(0.01,0.01) prior. For the coefficients η_{kl}, ($l = 0, 1, \ldots, r$) in the polynomial functions f_k, we specify $N(0, 100)$ prior distributions. In the time-to-event submodel, for Ψ_1 (and Ψ_2) and $\boldsymbol{\theta}_1$ (and also for $\boldsymbol{\theta}_2$), we specify multivariate normal prior with mean vector=0, and a diagonal covariance matrix with diagonal element=1000. We perform a sensitivity analysis, and the results (for some parameters) are summarized in Table 2.1. We notice that the hyperparameters have minimal effects on the final estimates.

We use MCMC iterations (based on Gibbs sampler and Metropolis-Hastings algorithm) for estimating the model parameters. We run 12,000 MCMC iterations, discard the first 2,000 as "burn-in", and use the remaining 10,000 iterations for estimating the model parameters. The estimated posterior densities and trace plots for some of the coefficients are provided in Figures 2.2-2.5. We note that these plots indicate a good convergence of the chains. In addition, we compute scale reduction factors (Brooks and Gelman, 1998 [12]) for assessing convergence of the chains, and all the computed scale reduction factors are smaller than 1.2, indicating a good convergence. We use sample means for estimating the respective model parameters. A 95% Bayesian credible interval is also computed for each coefficient based on the MCMC iterations.

TABLE 2.1: Results from sensitivity analysis for a set of coefficients in data analysis. IG, N, and Unif stand for the Inverse Gamma, Normal, and Uniform distributions, respectively.

Parameters	Prior distribution	Estimate
σ^2_{e1}	IG (0.01, 0.01)	0.47
-	IG (0.001, 0.001)	0.49
-	IG (1.2, 3.5)	0.43
σ^2_{e2}	IG (0.01, 0.01)	0.52
-	IG (0.001, 0.001)	0.55
-	IG (1.2, 3.5)	0.57
σ^2_{e3}	IG (0.01, 0.01)	0.38
-	IG (0.001, 0.001)	0.37
-	IG (1.2, 3.5)	0.41
Ψ_{11}	N(0,1000)	0.636
-	N(0,100)	0.633
-	Unif(-100,100)	0.638
Ψ_{12}	N(0,1000)	-0.418
-	N(0,100)	-0.414
-	Unif(-100,100)	-0.419
Ψ_{13}	N(0,1000)	-1.313
-	N(0,100)	-1.314
-	Unif(-100,100)	-1.311

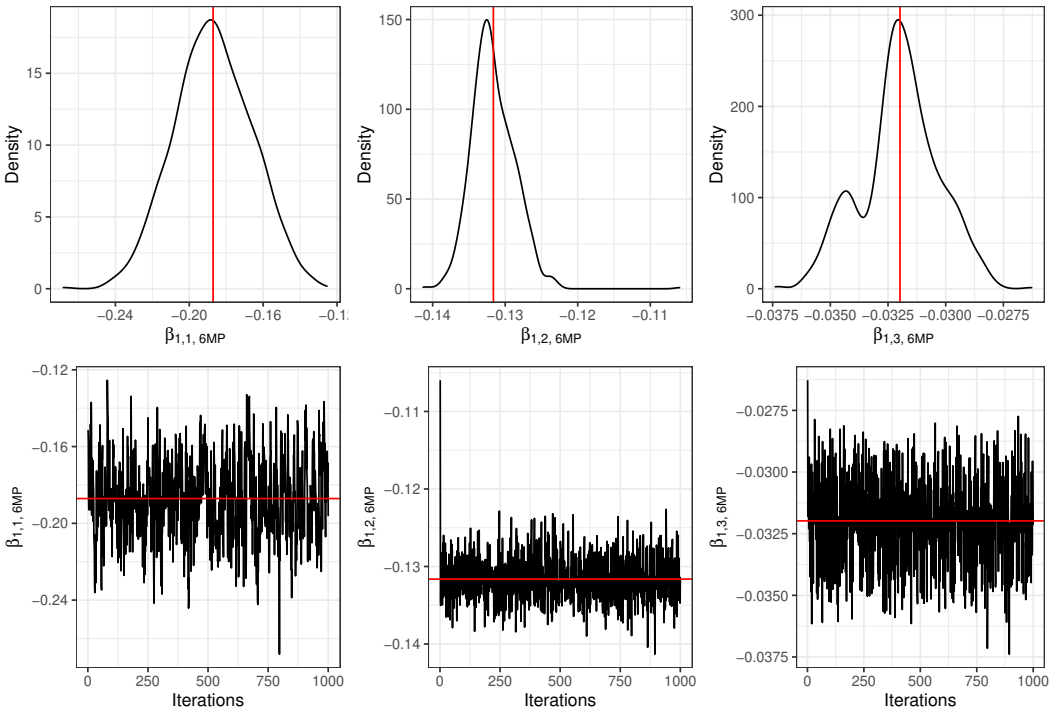


FIGURE 2.2: Estimated posterior density and trace plots for the fixed coefficients corresponding to the medicine 6MP in the data analysis.

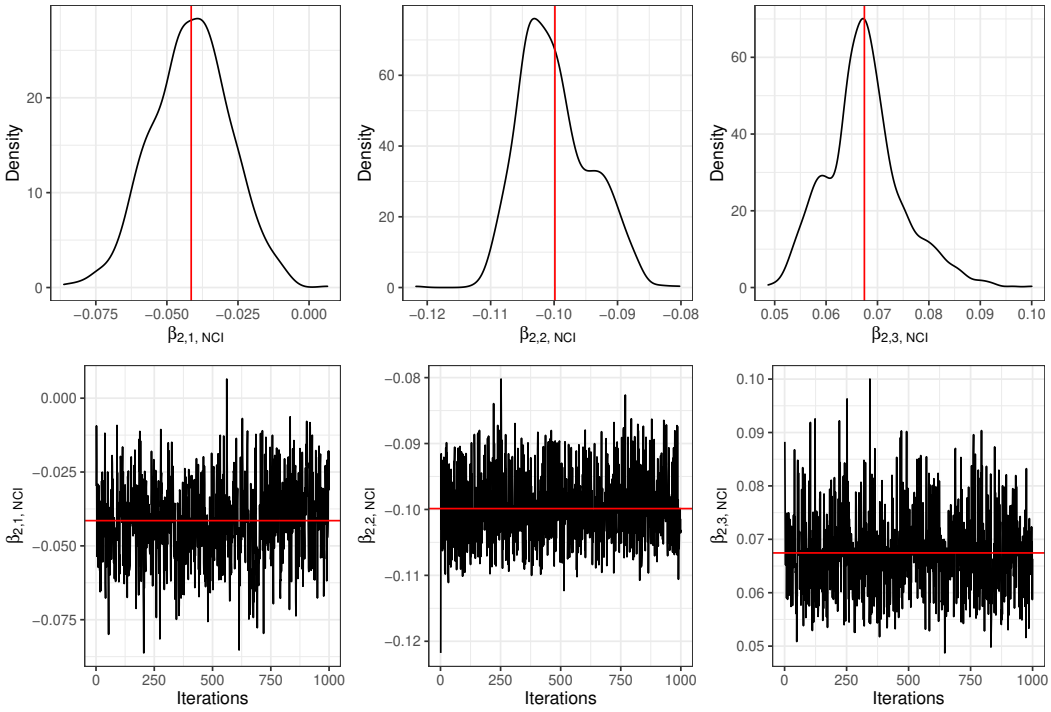


FIGURE 2.3: Estimated posterior density and trace plots for the fixed coefficients corresponding to the NCI risk for the longitudinal submodel in the data analysis.

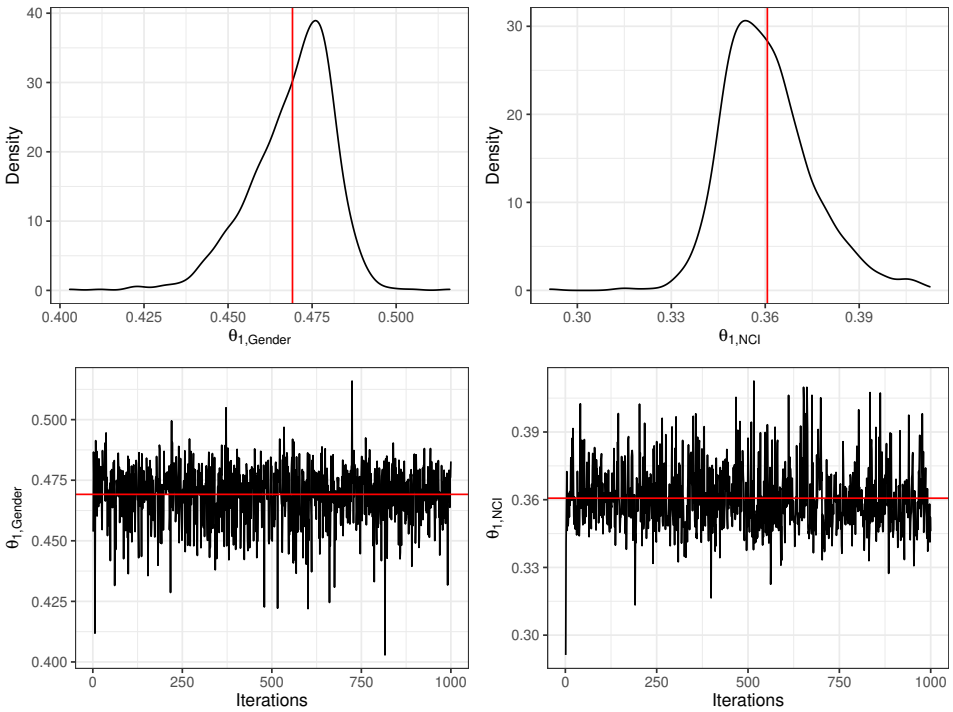


FIGURE 2.4: Estimated posterior density and trace plots for the fixed coefficients corresponding to gender and NCI risk for the time-to-event submodel in the data analysis.

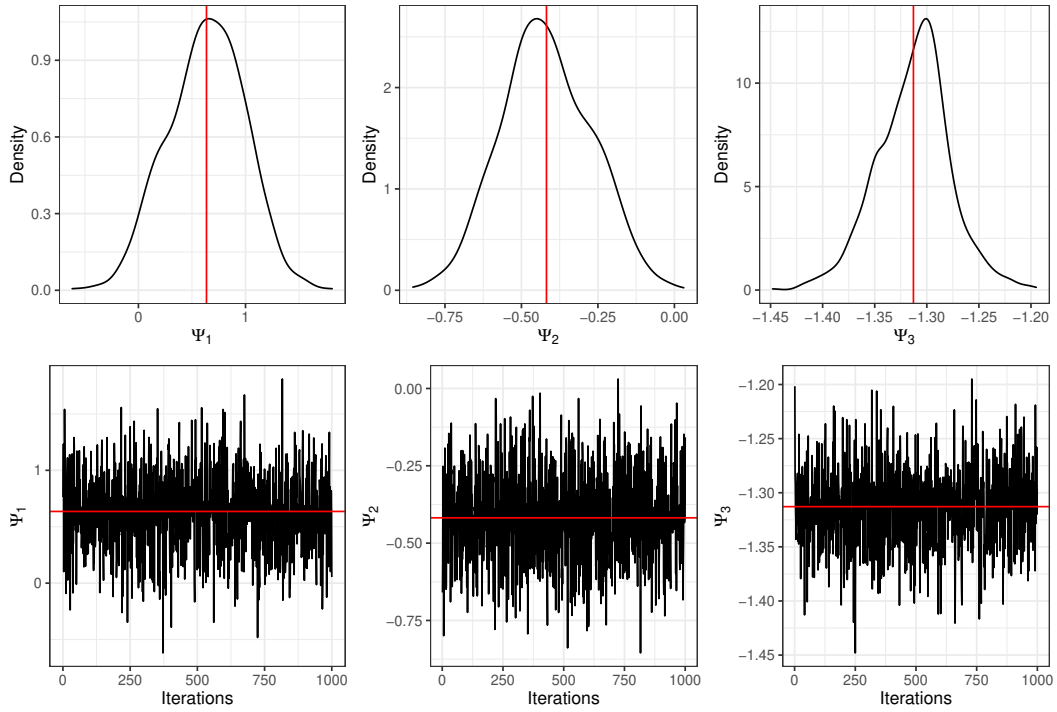


FIGURE 2.5: Estimated posterior density and trace plots for the three association parameters (Ψ) in the data analysis.

For obtaining the optimal order (r) of the polynomial functions f_k (in equation (2.1)) we consider the "complete DIC" (Celeux et al., 2006 [14]) which is computed based on the observed and the imputed missing observations (based on MCMC iterations), and is defined as DIC$=-4E\left[logf(Y,S|\boldsymbol{\alpha})\right]+2logf(Y,S|\hat{\boldsymbol{\alpha}})$. Note that $f(Y,S)$ denotes the joint density of the longitudinal and the event-time data as given in equation (2.4), and $\hat{\boldsymbol{\alpha}}$ is the estimated random effects. We consider only two choices, $r=2,3$; and we compute DIC values (shown in Table 2.2) for the four different specifications of the joint model as discussed in Section 2.3.2. The smallest (complete) DIC value is obtained for $r=2$ across all different joint models, and hence we consider r=2 for our analysis.

TABLE 2.2: DIC values for selecting the optimal order (r) of the polynomials f_k. Results are shown for the four different specifications of the joint model as discussed in Section 2.3.2.

	DIC			
r	Model I	Model II	Model III	Model IV
2	512.7	371.5	490.6	382.3
3	546.2	387.3	497.5	416.1

All our computations for this analysis are performed in R. In a Windows 10, i5 processor machine it takes nearly 24 hours for the complete analysis (including the posterior predictive inference in Section 2.3.4).

2.3.2 Model Selection

We compare the performance of several competing models, and select the one which provides the "best" fit and "best" prediction for our data. In the joint modeling literature, it is of great importance to evaluate the prediction accuracy and discriminative power of a model. Hence, we also compute the prediction error (expected error of predicting future events) and the area under the receiver operating characteristic curve (AUC) for different models under consideration. Note that AUC measures how effectively a joint model can discriminate the patients for which a relapse occurred from the patients with no relapse. Both there measures are computed within JMbayes package (Rizopoulos, 2016 [78]).

From equation (2.5) in Section 2.3.5, recall that $\pi_i(t + \Delta t|t)$ denotes the probability that the i-th patient will be event-free (no relapse) upto time $t + \Delta t$ given that it is event-free until time t. For a randomly chosen pair of patients $[i, j]$ who are event-free until time t, the discriminative power of a model is assessed by computing AUC as given below:

$$AUC = P\left[\pi_i(t + \Delta t|t) < \pi_j(t + \Delta t|t)|(T_i \in (t, t + \Delta t]) \cap (T_j > t + \Delta t)\right].$$

This means that in a fixed time interval $(t, t + \Delta t)$ if a relapse occurs for the i-th patient but the j-th patient is event-free upto time $t+\Delta t$, then the model must assign a higher non-relapse probability to the j-th patient. We use this criterion to compare different competing models along with the prediction error.

First, we consider random intercepts only in the longitudinal submodel, and use the expected longitudinal outcomes as predictors in the survival submodel (as shown in equation (2.2)). In other words, we specify $W_{ik}(t_{ij}) = a_{ik}$, where $\mathbf{a}_i = [a_{i1}, a_{i2}, a_{i3}]^T \sim N(0, D_1)$, as mentioned in Section 2.2.1, and specify the PH model given in equation (2.2) for the time-to-event. For the covariance matrix D_1, we consider an Inverse Wishart (4, M_1) prior, where M_1 is a diagonal matrix whose diagonal elements are generated from Gamma (0.5,0.01) distribution. We refer to this model as Model I.

Second, we consider random intercepts and random slopes of time in the longitudinal submodel. Specifically, we consider $W_{ik}(t_{ij}) = a_{ik} + b_{ik}t_{ij}$, where $[\mathbf{a}_i^T, \mathbf{b}_i^T]^T \sim N(0, D_2)$ (refer to Section 2.2.1), and the model given in equation (2.2) is used for the time-to-event. For the covariance matrix D_2, we consider an Inverse Wishart (7, M_2) prior, where M_2 is a diagonal matrix whose diagonal elements are generated from Gamma (0.5,0.01) distribution. We refer to this model as Model II.

Third, we specify $W_{ik}(t_{ij}) = a_{ik}$ in the longitudinal submodel given in equation (2.1), and consider the PH model given in equation (2.3) for the time-to-event. We specify $W_i^*(t) = c_i$, and assume that the vector of random effects $[a_{i1}, a_{i2}, a_{i3}, c_i]^T \sim N(0, D_3)$. Note that the association between the time-to-event and the biomarkers is captured by the covariance matrix D_3. For D_3 we consider an Inverse Wishart prior similar to Models I and II. We refer to this model as Model III.

Finally, we specify $W_{ik}(t_{ij}) = a_{ik} + b_{ik}t_{ij}$, in the longitudinal submodel given in equation (2.1), and consider the PH model given in equation (2.3) for the time-to-event assuming $W_i^*(t) = c_i + d_i t$. Thus, we consider patient-specific random intercepts and random slopes of time in the PH model. Further, we assume that $[\mathbf{a}_i^T, \mathbf{b}_i^T, c_i, d_i]^T \sim N(0, D_4)$, and the covariance matrix D_4 captures the association between the biomarkers and the time-to-event. This model is referred to as Model IV.

Table 2.3 shows prediction error (PE), and AUC values for the four competing models (with r=2 for all specifications) for t=100 (duration of treatment) and for three different choices of Δt. We notice that Model II provides higher AUC values and lower PE values than the other three models, consistently. Hence, we select Model II as the "best model" for our dataset.

Next, we fit Model II to our data and compute the DIC as discussed in Section 2.3.1. Additionally, we fit two separate models. For the longitudinal biomarkers we fit the linear mixed models given in equation (2.1), and for time-to-event we use the PH model (similar to equation (2.2)) with the biomarkers (observed) and the fixed covariates as predictors. We compute the DIC for these two models, and add them to get the combined DIC. The DIC value for the joint model is 371.5, where the combined DIC for the separate modeling is 493.8. Thus, we conclude that the proposed joint model gives a better fit to our data.

TABLE 2.3: Model selection in ALL data analysis. Prediction Error (PE), and AUC values (for t=100, and $\Delta t = 50, 100, 150$) are given for the four competing models, described in Section 2.3.2.

	Model I	Model II	Model III	Model IV
AUC(t=100,Δt=50)	0.33	0.71	0.43	0.44
AUC(t=100,Δt=100)	0.35	0.72	0.45	0.48
AUC(t=100,Δt=150)	0.37	0.75	0.51	0.55
PE(t=100,Δt=50)	0.11	0.04	0.09	0.07
PE(t=100,Δt=100)	0.13	0.06	0.10	0.08
PE(t=100,Δt=150)	0.15	0.08	0.11	0.10

2.3.3 Findings

In Table 2.4, we summarize the estimated coefficients and 95% Bayesian credible intervals (based on MCMC iterations) for the longitudinal submodel. We consider a covariate as "significant" if the corresponding estimated 95% CI does not contain a zero (Das, 2016 [19]).

We notice that two medicines are significant for all the three biomarkers. We also note that the effects of 6MP are negative and effects of MTx are positive for all the three biomarkers. The NCI risk, presence of bulky disease, presence of CNS disease, morphological remission, and Day 35 risk are significant for all the three biomarkers. Age and gender are significant for ANC and platelet count. Risk at presentation is

significant only for WBC count. Lineage (T/B cells) and risk at day 8 are significant for WBC count and platelet count. MRD status is significant only for platelet count.

In Table 2.5, we summarize the effects of the time-invariant covariates, and the biomarkers on the relapse time. Interestingly, we notice that ANC and platelet count are significant for the relapse time but WBC count is not significant. Additionally, the effects of ANC and platelet count are negative, indicating that a higher (mean) ANC and a higher (mean) platelet count will result in a reduced hazard (i.e. a lower probability of relapse). Note that in Table 2.4 we noticed that 6MP has negative effects, and MTx has positive effects on the biomarkers. Combining the findings, we infer that a lower dose of 6MP and a higher dose of MTx can be recommended for reducing the relapse probability. Among the other covariates, except age, all the other covariates are significant for the relapse time. Lineage, risk at day 8, MRD status, and morphological remission have negative effects, and all the other covariates have positive effects.

In Figure 2.6, we plot the (time-varying) correlations among the biomarkers. We notice that the correlation between ANC and the platelet count is mostly higher than the other two correlations. However, the correlation between any two biomarkers decrease over time during treatment, possibly due to the effects of the medicine.

TABLE 2.4: Estimated coefficients and corresponding 95% CIs for the covariates in the longitudinal submodel in ALL data analysis.

	WBC count		Neutrophil count		Platelet count	
Covariate	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
6MP dose	-0.187	(-0.227,-0.147)	-0.132	(-0.137,-0.126)	-0.032	(-0.035,-0.029)
MTx dose	0.075	(0.033,0.119)	0.065	(0.058,0.071)	0.015	(0.013,0.018)
Age at diagnosis	0	(-0.018,0.018)	-0.016	(-0.022,-0.01)	-0.111	(-0.117,-0.102)
WBC at presentation	0.003	(-0.005,0.01)	-0.003	(-0.011,0.001)	-0.018	(-0.022,-0.014)
Gender	-0.012	(-0.03,0.008)	-0.04	(-0.047,-0.031)	-0.052	(-0.059,-0.041)
Lineage	0.07	(0.04,0.101)	0	(-0.013,0.021)	-0.036	(-0.059,-0.022)
NCI risk group	-0.041	(-0.068,-0.015)	-0.1	(-0.109,-0.088)	0.067	(0.054,0.085)
Bulky disease	-0.059	(-0.084,-0.038)	-0.043	(-0.058,-0.032)	-0.053	(-0.07,-0.043)
CNS disease	-0.039	(-0.059,-0.015)	0.126	(0.114,0.147)	-0.058	(-0.065,-0.045)
Risk at presentation	0.039	(0.021,0.057)	0.011	(-0.001,0.035)	-0.019	(-0.028,0.001)
Day 8 risk	0.053	(0.032,0.074)	0.01	(-0.014,0.026)	-0.028	(-0.042,-0.017)
Day 35 risk	-0.06	(-0.071,-0.049)	-0.039	(-0.044,-0.028)	0.013	(0.007,0.019)
Morphological remission	0.083	(0.049,0.125)	0.084	(0.069,0.127)	0.186	(0.172,0.209)
MRD status	0.001	(-0.013,0.014)	0.01	(-0.004,0.017)	-0.075	(-0.084,-0.071)

TABLE 2.5: Estimated coefficients and corresponding 95% CIs for the covariates in the time-to-event submodel in ALL data analysis.

Covariate	Estimate	95% CI
WBC count	0.636	(-0.075,1.29)
Neutrophil count	-0.418	(-0.695,-0.131)
Platelet count	-1.313	(-1.387,-1.241)
Age at diagnosis	0.013	(-0.012,0.043)
WBC at presentation	-0.033	(-0.051,-0.015)
Gender	0.469	(0.443,0.488)
Lineage	-0.531	(-0.606,-0.465)
NCI risk group	0.361	(0.338,0.393)
Bulky disease	0.07	(0.026,0.102)
CNS disease	0.375	(0.314,0.432)
Risk at presentation	0.297	(0.279,0.321)
Day 8 risk	-0.269	(-0.307,-0.234)
Day 35 risk	0.364	(0.352,0.381)
Morphological remission	-0.841	(-0.896,-0.771)
MRD status	-0.246	(-0.267,-0.231)

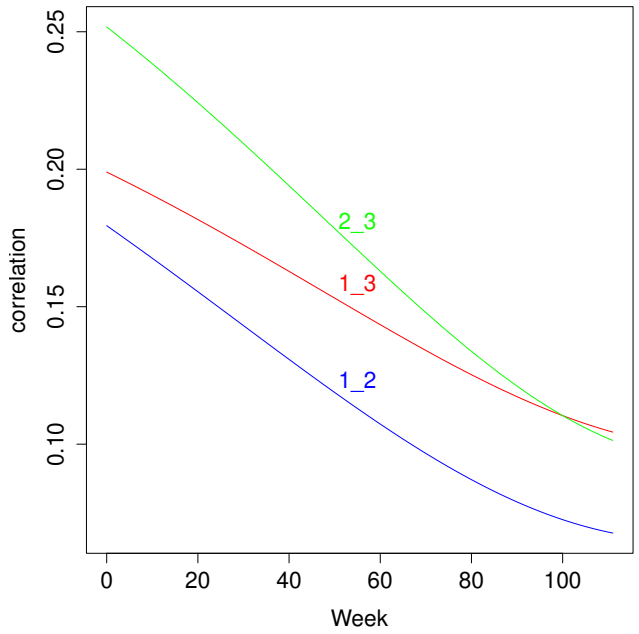


FIGURE 2.6: Estimated time-varying correlations among the three biomarkers. Biomarkers 1, 2 and 3, respectively, refer to WBC count, neutrophil count (ANC) and platelet count.

2.3.4 Posterior Predictive Inference

Next, we investigate the specific effects of some significant covariates on the longitudinal and the event time outcomes through posterior predictive distributions. Posterior predictive checks are used for assessing the model fit with missing and latent data (Gelman et al. 2005 [34]). Let (Y_i^{rep}, S_i^{rep}) and (Y_i^{obs}, S_i^{obs}), respectively, denote the replicated and the observed datasets for both the longitudinal and event time outcomes, and let $\boldsymbol{\Theta}$ denote the set of all model parameters (including random effects) in the proposed joint model. The posterior predictive distribution of the replicated data conditional on the observed dataset is given as follows:

$$P(Y_i^{rep}, S_i^{rep}|Y_i^{obs}, S_i^{obs}) = \int P(Y_i^{rep}, S_i^{rep}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|Y_i^{obs}, S_i^{obs})d\boldsymbol{\Theta}. \tag{2.9}$$

For our analysis, we first fix a specific covariate of interest (for example, fix gender as male), and then sample the other covariates from their respective empirical distributions (based on the given dataset). Model parameters are sampled from their respective posterior densities. We sample data for 50,000 patients over 277 time points, where the biomarkers are measured for the first 30 time points (the average number of visits in the actual dataset). However, we consider a longer follow-up period (than the data at hand) for a better prediction of relapse.

In Figure 2.7, we show the mean predicted (log transformed) ANC, and the mean predicted platelet count for two genders. We also show the predicted (median) non-relapse probabilities in this figure. We notice that the mean ANC and the mean platelet count for females are consistently higher than those for males. The jumps (for both these plots) at week 30 is due the fact that no treatment is given after that week. Additionally, we observe that the non-relapse probabilities are higher for females over time. This is consistent with the results shown in Table 2.5, where we saw that the effects of (mean) ANC and (mean) platelet count are negative on hazard.

Next, we investigate the effect of NCI risk group on the biomarkers and the relapse time. In Figure 2.8, we notice that ANC for "high-risk" (HR) group is uniformly higher than "standard-risk" (SR) group, but the trend is the reverse for the platelet count. Additionally, we see that the non-relapse probability for the HR group is higher than the SR group. It is possibly because the patients in the HR group (at presentation) are treated more carefully than those in the SR group. But this is indeed a very interesting finding from our analysis.

Next, we focus on the two drugs which are found to be significant for the biomarkers. The doses are grouped into high, medium, and low; based on the 0.75 quantile and the 0.40 quantile of their respective empirical distributions. We consider all the nine different dose combinations (e.g. high-high, medium-low etc.), and plot the estimated mean curves for each group. The dose combination "high-low" was rarely given (in the actual dataset), and hence we do not include it here. If a particular

dose combination is given to a particular patient only for more than 15 weeks (i.e. more than 50% of the times), then we assign the patient to that specific dose group. The patients who are not treated with one specific dose combination at least for 15 weeks are discarded, and we end up getting nearly 38,000 patients who are given a particular dose combination at least for 15 weeks. In Figure 2.9, we notice that a low dose of 6MP and a high dose of MTx result in the highest ANC across the weeks. On the other hand, high doses for both the medicines result in the lowest ANC. For the platelet count, we notice a decreasing trend for all the groups. The curve for "low-medium" group is consistently higher while that for "high-high" group is consistently lower than the other groups. In Figure 2.9, it is noted that, in general, higher ANC and higher platelet counts are observed for low (or medium) dose of 6MP and high (or medium) dose of MTx. Thus, on the average, we recommend a lower dose of 6MP and a higher dose of MTx during treatment.

Finally, we focus on the risk-at-presentation which was found to be significant for the relapse time (but not significant for ANC and platelet counts). There are three groups, i.e. high risk (HR), standard risk (SR), and intermediate risk (IR) groups. In Figure 2.10, we observe that the (median) non-relapse probability is higher for the HR group, and it is the lowest for the IR group. A similar trend was observed for NCI risk groups (in Figure 2.8), and this again reflects that the patients assigned to the HR group (at presentation) show higher non-relapse probabilities in the follow-up period. Rhein et al. (2011) [75] reported a similar inference based on their analysis on ALL patients. This interesting (and counter-intuitive) inference needs further investigation.

2.3.5 Subject-wise prediction of relapse probability

We compute the relapse probability for each patient (for which no relapse is observed during treatment) at different time points in the follow-up period. This is computed using *JMbayes* package in *R*. In Figure 2.11, we show the (predicted) relapse probabilities for some randomly selected patients. For each patient, the trajectory starts when the treatment (for that patient) ends. Conditional on the data (on biomarkers and the predictors) available until the end of the treatment, we compute the relapse probabilities for each patient. The relapse probability increases over time, as expected, and we group the patients based on their maximum (predicted) relapse probability corresponding to the last time point in the follow-up period. For the "low-risk" group, the predicted relapse probabilities are all smaller than 0.10. For the "moderate-risk" group, those are between 0.10 and 0.40. All the other patients belong to the "high-risk" group. The thresholds (0.10 and 0.40) are chosen based on doctors' recommendation. We note that for 85% of the children belonging to the "high-risk" group relapse occurred (during follow-up) in the actual dataset. For the "moderate risk" group and the "low-risk" group, the occurrences are 25% and 7%, respectively, indicating a good predictive power of our model.

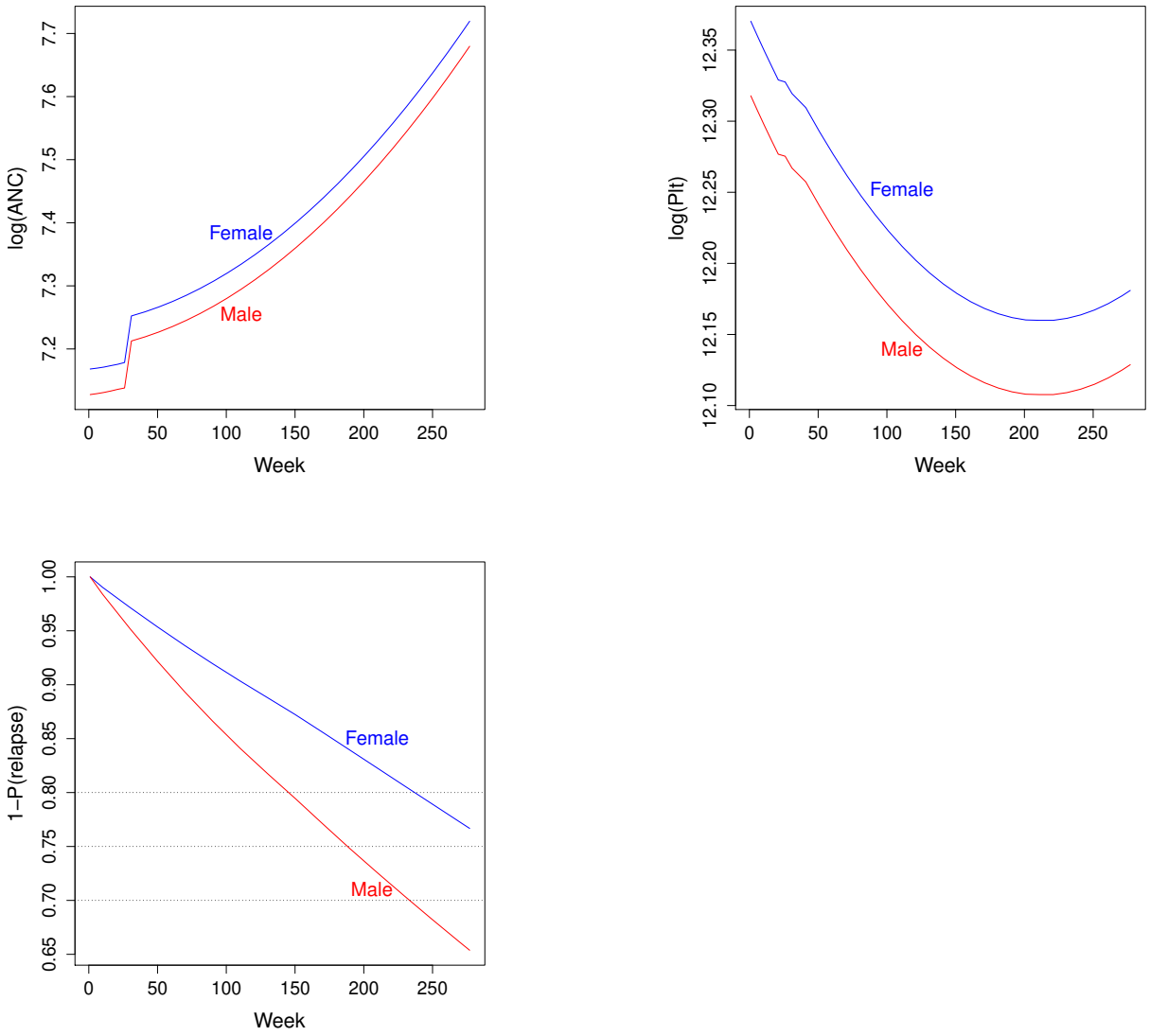


FIGURE 2.7: Estimated mean curves for the neutrophil count, platelet count; and the estimated (median) non-relapse probabilities for the two genders (M/F) in ALL data analysis.

2.4 Simulation Study

We perform a simulation study for assessing the effectiveness of Bayesian imputation of the missing responses in joint modeling. We simulate a dataset which is quite similar to the ALL dataset. We consider three biomarkers measured from a set of 200 subjects over twenty evenly spaced time points. We consider ten covariates, two of them are time-varying. Biomarkers are simulated using equation (2.1). At time T=20, the treatment phase is over, and the follow-up period starts. The details for the simulation of the covariates and the longitudinal biomarkers are as follows:

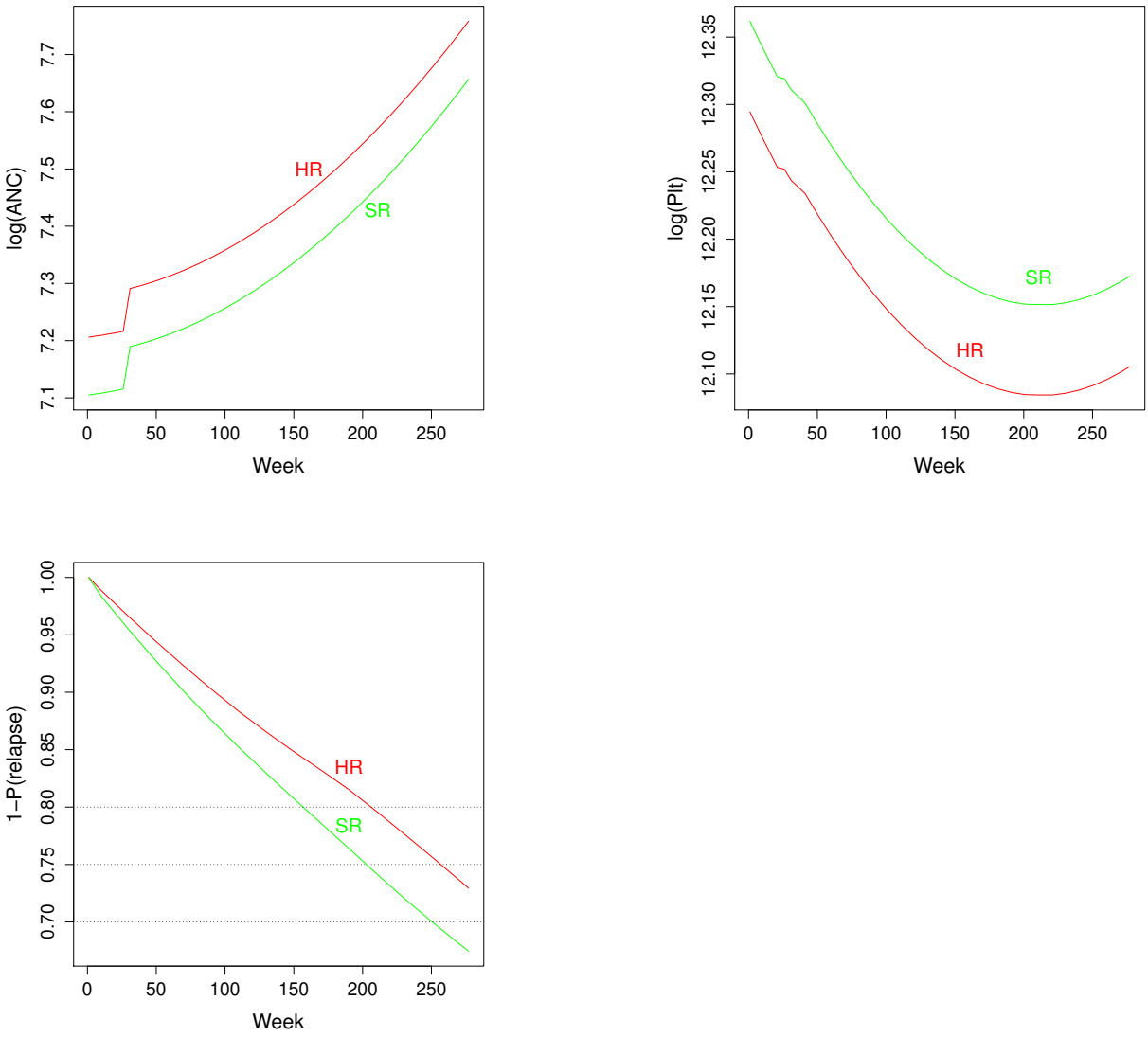


FIGURE 2.8: Estimated mean curves for the neutrophil count, platelet count; and the estimated (median) non-relapse probabilities for the two NCI risk groups [High-Risk (HR), and Standard-Risk (SR)] in ALL data analysis.

The time-varying covariates, X_1 and X_2, are simulated from the following auto-regressive model:

$$X_{it} = \alpha X_{i,t-1} + \epsilon_{it}, \tag{2.10}$$

where ϵ_{it} are iid N(0,1), and α=0.85. For the first time point (t=1) we sample from a standard normal distribution, and then use the above model for generating the covariates. Among the eight fixed covariates, we consider four as binary, and they are generated independently from a Bernoulli distribution with success probability=0.48. The other four covariates are generated, respectively, from N(0,10), Gamma(1,3), Uniform (2,5), and Beta (1.5, 2.5) distributions independently. The biomarkers are sampled by using the model given in equation (1) of the main document with the

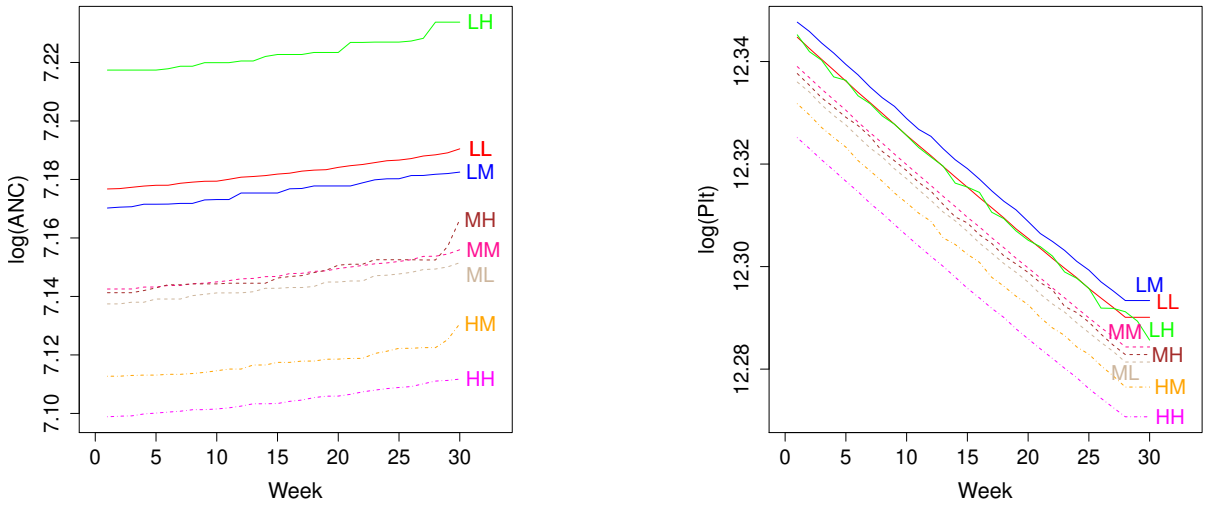


FIGURE 2.9: Estimated mean curves for the neutrophil count and platelet count for different medicine doses in ALL data analysis. The curve with label (ij) refers to the i-th level of 6MP and the j-th level of MTx, where i, j=high (H), medium (M) or low (L).

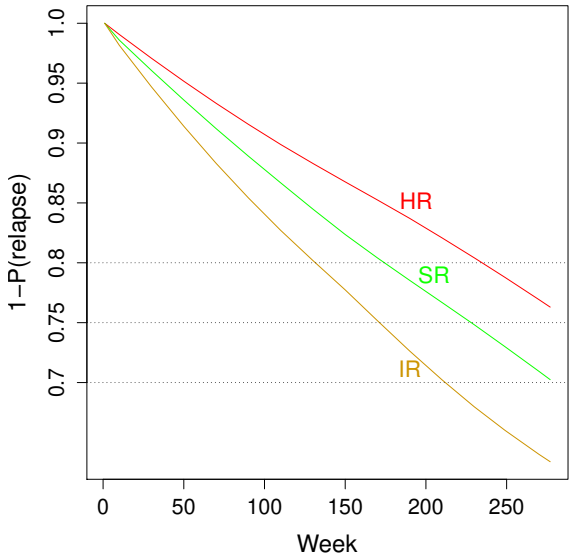


FIGURE 2.10: Estimated (median) non-relapse probabilities for the three risk groups at presentation [High-Risk (HR), Standard-Risk (SR), and Intermediate-Risk (IR)] in ALL data analysis.

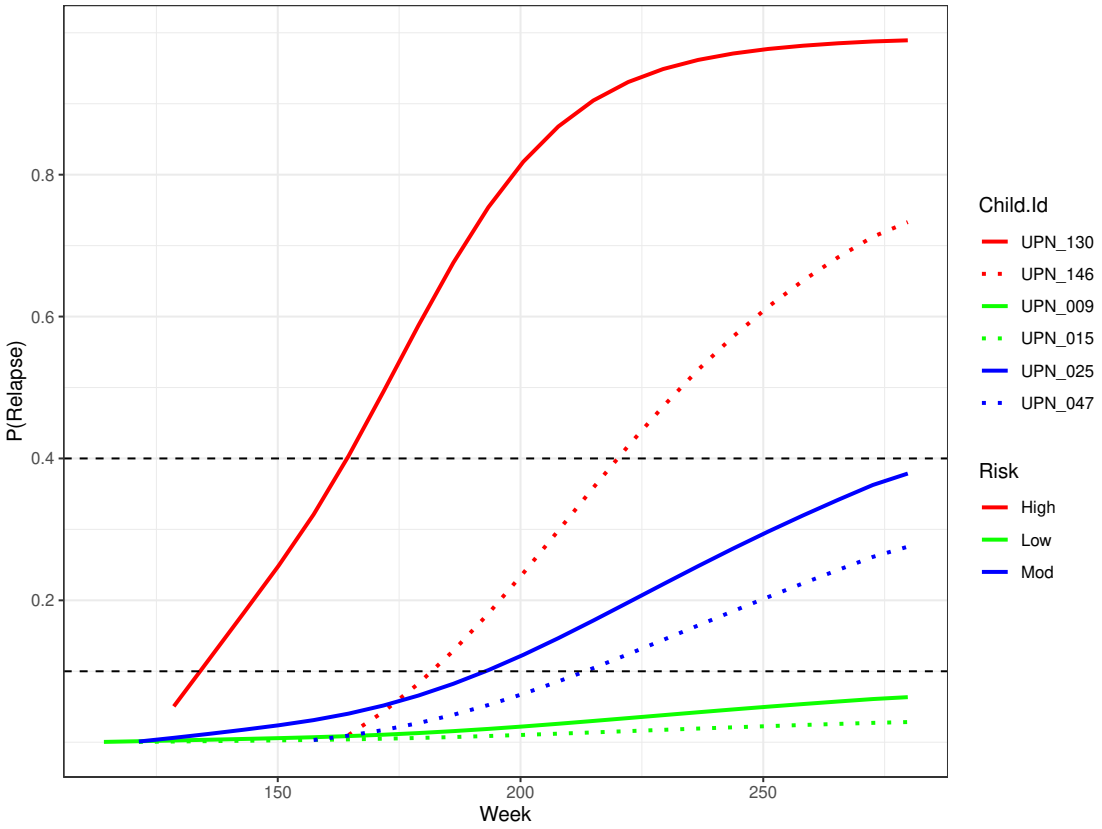


FIGURE 2.11: Subject-wise predicted relapse probabilities for some randomly selected patients in ALL dataset.

specification $f_k(t) = 1.5 + 2.3t$. The residuals are independently generated from N(0,1) distribution.

We consider, $\beta_{11} = [0.95, 1.36]$; $\beta_{12} = [-1.45, 0.86]$; $\beta_{13} = [1.07, 0.75]$.
$\beta_{21} = [-0.34, 0.93, -1.23, 1.45, 2.34, -0.57, 0.84, 1.05]$;
$\beta_{22} = [0.64, -0.73, -1.03, 2.15, 1.34, -0.63, 0.24, 0.65]$;
$\beta_{23} = [0.78, -0.57, -1.34, 1.05, 1.21, -0.61, 0.39, 0.88]$.

Finally, for simulating the event time using the PH model given in equation (2) of the main document, we use $\Psi_{11} = -0.15$, $\Psi_{12} = -1.46$, $\Psi_{13} = -2.09$; $\theta_{11} = 0.05$, $\theta_{12} = -0.68$, $\theta_{13} = 1.15$, $\theta_{14} = -2.33$, $\theta_{15} = 0.21$, $\theta_{16} = -0.29$, $\theta_{17} = 1.36$, $\theta_{18} = 1.89$. These values are chosen based on some existing literature and also in the same line as the estimates from ALL data analysis.

We simulate time-to-event for the subjects assuming that the subjects are censored at T=50, when the study ends. Thus, the patients are followed for the next 30 time points after the end of treatment. We use equation (2.2) for generating the time-to-event, the values of the model parameters as mentioned above. Once the complete dataset is generated, we randomly create some missing values in the three biomarkers. We consider 32% missing values in one biomarker, 5% missing values in the second biomarker, and 4% missing values in the third one (similar to the ALL

dataset). For each biomarker, we first randomly select some subjects for which we create some missing observations again at some randomly selected time points. We use *missMethods* library in *R* for creating missing biomarkers in a complete dataset.

We consider three alternative approaches. First, we consider only the subjects with no missing values, and we refer to this approach as "Completers only". Second, we consider the available dataset where the missing observations are treated as missing values and we do not impute those. This approach is referred to as "Available data only". Finally, we consider the Bayesian imputation using the joint model. In all the three approaches, the model parameters are estimated using 10,000 MCMC iterations.

Next, based on 200 replications we compute the average absolute bias and average width of the 95% Bayesian credible intervals (and the coverage probabilities) of the regression coefficients. In Table 2.6 we show the results for a selected set of coefficients. We notice that the "Completers only" approach performs the worst since it results in a larger average bias and a wider credible interval almost for all the coefficients. Bayesian imputation of the missing biomarkers using the joint model is worth since this approach results in the smallest average bias, and the shortest credible intervals with reasonable coverage probabilities.

TABLE 2.6: Average absolute bias, width of 95% CIs, and the estimated coverage probability (C.P.) for a set of regression coefficients for the three competing approaches in the simulation study.

	Completers only		Available data only		Bayesian Imputation	
Coefficient	Bias	width(C.P.)	Bias	width (C.P.)	Bias	width (C.P.)
β_{111}	0.93	2.42(0.96)	0.34	1.36(0.95)	0.11	1.02(0.95)
β_{112}	0.88	2.17(0.96)	0.26	1.15(0.95)	0.13	0.97(0.94)
ψ_{11}	0.76	1.03(0.96)	0.32	0.98(0.95)	0.08	0.83(0.95)
ψ_{12}	0.79	1.18(0.95)	0.38	1.01(0.95)	0.11	0.78(0.95)
θ_{11}	0.81	2.12(0.95)	0.55	1.21(0.96)	0.10	0.92(0.95)
θ_{12}	0.86	2.04(0.95)	0.49	1.16(0.95)	0.12	1.10(0.94)

2.5 Summary

Despite the significant improvement in its survival rate over the years, ALL is still globally considered as the main cause of death from cancer among children. ALL typically occurs more often in Caucasians, Hispanics, and Latin Americans than in Africans (Renbarger et al., 2008 [74]), and therefore, there exists a vast literature on ALL for the United States and for the European countries. However, there is a lack of similar significant work for India, and in general for most of the Asian and the African countries. In this chapter we present a comprehensive study for the Indian children (diagnosed as ALL patients) and address some interesting research questions. We develop a Bayesian joint model which can impute the missing longitudinal data within

each iteration of MCMC, and also can dynamically predict the non-relapse probabilities for each subject.

We note that for the analysis presented in this chapter we used Bayesian hierarchical models for our joint modeling. Different parts of the model are proposed in the literature before but our analysis needs to combine those systematically. However, the main novelty of our approach is its ability to impute the missing outcomes within each iteration of MCMC by appropriately considering the dependence structure within and between the outcomes. Also, since we have limited patients in the clinical trial we perform posterior predictive checks for consistent inference.

However, we note that our current analysis ignores the genetic factors associated with the development and progression of ALL since such information was not available in the given dataset. Additionally, we assess the effects of the covariates on the mean longitudinal outcomes, and also the effects of the mean longitudinal outcomes on the event-time. Assessing the effects of different covariates at different quantile levels (instead of the mean) of the biomarkers provides a better understanding of the complex association among the biomarkers, covariates and event-time. Specifically, when the joint distribution of the biomarkers deviates from a multivariate normal distribution quantile-based analysis is more meaningful due to its robustness. In Chapter 3 we develop a quantile-based joint model for ALL dataset.

Chapter 3

A Bayesian quantile joint modeling for multivariate longitudinal and event-time data

3.1 Preamble

In Chapter 2, we present a Bayesian joint model for simultaneously modeling the progression of the longitudinal biomarkers, and the time-to-event. While such models are quite effective, and provide some interesting insights they are based on the multivariate Gaussian assumption for the joint distribution of the biomarkers. In addition, such models can model the evolution of the mean biomarkers, and can assess the effects of the mean biomarkers on the event-time. In many real applications, the biomarkers are non-Gaussian and/or the goal is to assess the effects of different quantiles of the biomarkers on the event-time. The model we proposed in Chapter 2 fails to work under this setting.

Quantile regression (QR) model, originally developed by Koenker and Bassett (1978) [49], models and predicts the quantiles of the outcome(s) and can assess the effects of the covariates on the outcome(s) at different quantile levels. A Bayesian version of the quantile regression model was first proposed in Yu and Moyeed (2001) [103] who considered an Asymmetric Laplace Distribution (ALD) for the outcome. Koenker (2004) [48], Geraci and Bottai (2007) [35] developed QR models for longitudinal outcomes. Kozumi and Kobayashi (2011) [50] showed that an ALD can be expressed as a mixture of a normal and an exponential distribution, and they developed an efficient Gibbs sampler for estimating the regression coefficients. Although the literature on QR is quite rich, there are limited works on QR for the joint modeling of longitudinal and event-time data. Farcomeni and Viviani (2015) [28] first proposed a linear quantile mixed model for such joint modeling, and they developed a Monte Carlo Expectation Maximization (MCEM) algorithm for estimating the model parameters. More recently, Yang et al. (2019) [101] developed a Bayesian quantile regression joint model to predict the risk of developing Huntington's disease. Their

model can dynamically predict the subject-specific survival probabilities, and hence can provide interesting clinical insights. Zhang and Huang (2020) [107] developed a quantile regression based Bayesian joint model to analyze the Multicenter AIDS Cohort Study data. However, they consider a univariate biomarker and developed the computational algorithm for jointly modeling a longitudinal biomarker and an event-time at different quantile levels of the biomarker.

In our application, we note that the neutrophil count is indeed a part of WBC count, and the evolution of ALL is heavily influenced by the lymphocyte count which is another major part of the WBC count. Therefore, instead of modeling WBC, neutrophil count and platelet count we focus on the joint modeling of lymphocyte count, neutrophil count and platelet count. Similar to Chapter 2, the now develop Bayesian joint model for analysing the biomarkers and the event-time at different quantile levels of the biomarkers. We extend the work in Zhang and Huang (2020) [107] for a multivariate setting, and consider a linear quantile mixed model for modeling three longitudinal biomarkers (Kulkarni et al., 2019 [51]). The longitudinal dependence and the dependence among the biomarkers are modeled by subject-specific random effects for which we consider a multivariate Brownian motion for higher flexibility (Picchini et al., 2010 [65]). We exploit the mixture representation of ALD (Biswas and Das, 2021 [9]) for the computational ease, and develop an efficient Gibbs sampler algorithm for our computation. For modeling the relapse-time, we use a semi-parametric proportional hazards (PH) model where the baseline hazard function is modeled using a B-spline (Rizopoulos, 2016 [78]). Our analysis shows that a relapse is accelerated by a higher lymphocyte count. We also notice that the effects of the fixed covariates on the outcomes differ across quantiles. Across all quantiles 6MP reduces the lymphocyte counts, and MTx increases the neutrophil counts. Both the drugs control the platelet count but the effects vary from one quantile level to the other.

We compute the median estimated non-relapse probabilities for different quantile levels, and also plot the estimated quantiles for each biomarker separately. We do not come across a quantile crossing issue in our analysis. Additionally, we compute the covariance structure for the three biomarkers at different quantile levels.

The rest of this Chapter is organized as follows. In Section 3.2, we discuss the dataset in detail, and also discuss the motivation for our quantile-based joint modeling. In Section 3.3, we describe the proposed model and the joint posterior distribution. The findings from the data analysis are discussed in Section 3.4. In Section 3.5, we perform a simulation study for evaluating the predictive power of the proposed model. Finally Section 3.6 concludes.

3.2 ALL Chemotherapy Dataset and Motivation

Our dataset for this analysis is the same as discussed in Section 1.4.1. However, since neutrophil count is a part of WBC, and the progression of leukemia is majorly controlled by the lymphocyte count (note that the disease is named as acute lymphocytic leukemia), we consider the lymphocyte count as one of the biomarkers instead of WBC. All the other variables are exactly the same as in Chapter 2. Table 3.1 provides the list of all variables and their roles in our analysis. The summary statistics of the fixed covariates are provided in Table 1.3 (in the Chapter 1).

TABLE 3.1: Variables used in the ALL data analysis. The role of each variable in our model is also specified.

Name	Type	Role (in our model)
Lymphocyte count	Continuous (longitudinal)	Outcome
Neutrophil count	Continuous (longitudinal)	Outcome
Platelet count	Continuous (longitudinal)	Outcome
Relapse time	time-to-event	Outcome
Dose of 6MP	Continuous (longitudinal)	covariate
Dose of MTx	Continuous (longitudinal)	covariate
Age at diagnosis	Continuous	Fixed Covariate
WBC count at presentation	Continuous	Fixed Covariate
Gender	Binary	Fixed Covariate
Lineage	Categorical	Fixed Covariate
NCI risk group	Categorical	Fixed Covariate
Bulky disease	Binary	Fixed Covariate
CNS disease	Binary	Fixed Covariate
Risk at presentation	Categorical	Fixed Covariate
Day 8 risk	Categorical	Fixed Covariate
Day 35 risk	Categorical	Fixed Covariate
Morphological Remission	Categorical	Fixed Covariate
MRD status	Categorical	Fixed Covariate

We consider the lymphocyte count, neutrophil count (ANC) and platelet count as our longitudinal biomarkers, and for stabilizing the variances we use the log transformed values. Figure 3.1 shows the log transformed longitudinal trajectories for three outcomes. The plot indicates that the log transformation stabilizes the variances of all the three biomarkers under consideration.

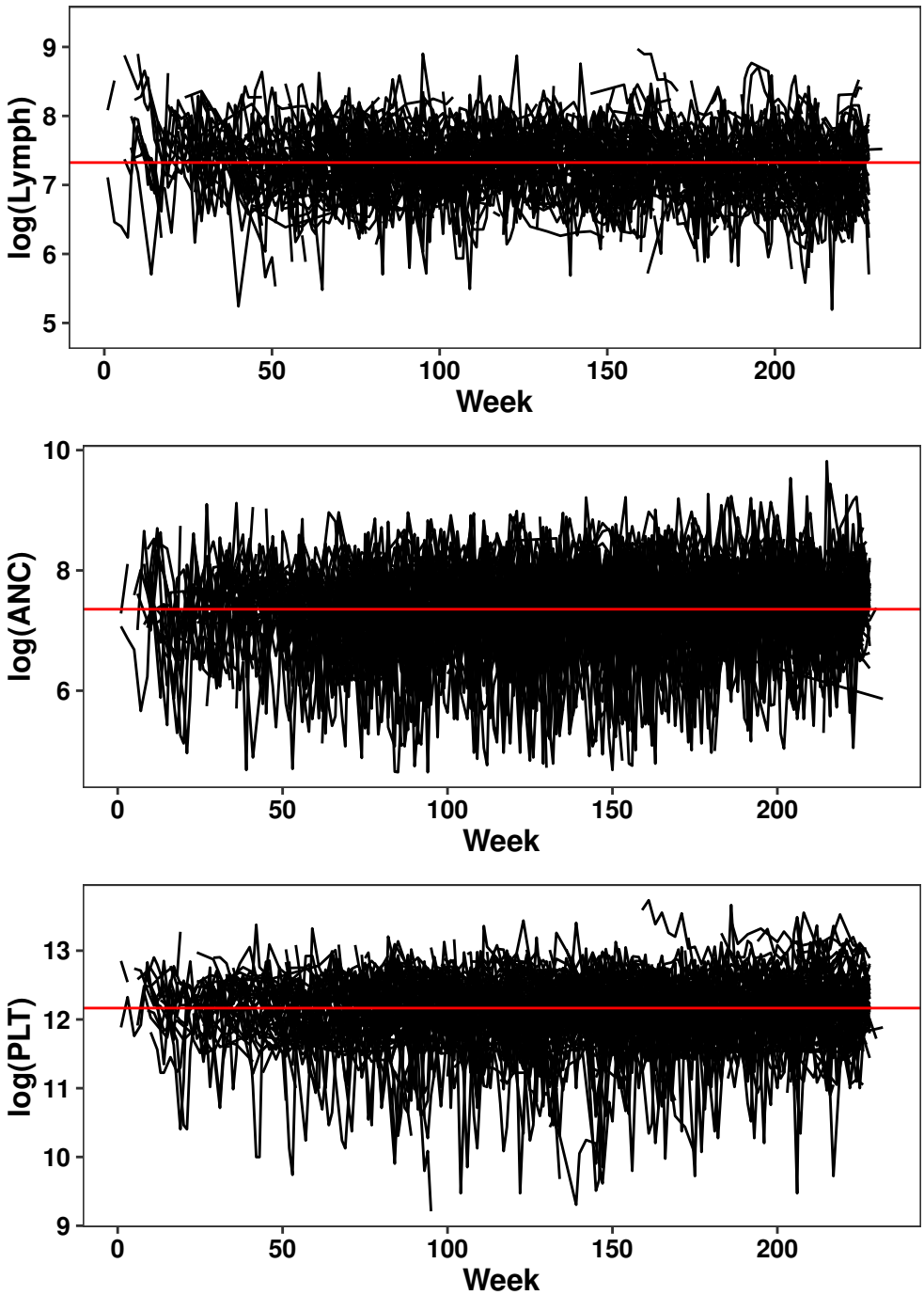


FIGURE 3.1: Longitudinal (log-transformed) biomarkers (Lymphocyte count, ANC, and Platelet count) in the ALL dataset.

In Figure 3.2 we show the longitudinal (raw) trajectories of the three biomarkers for four randomly selected patients. The solid and the dotted curves are used, respectively, for the patients with an event (relapse) and with no event. The solid horizontal lines denote the average biomarker values. We observe a higher (than the average) lymphocyte count, and a lower (than the average) ANC and a lower platelet count for the patients for which the event occurred. This indicates that the event-time is

possibly influenced by the observed biomarker values. Hence the research goal is (i) to study the progression of biomarkers, (ii) to study the effects of the biomarkers on the event-time, and (iii) effects of the covariates on the longitudinal process (for the biomarkers) and on the event-time. Following Kundu et al. (2023) [52] we go for a joint modeling of the biomarkers and event-time for powerful Statistical inference.

In Figure 3.3 we show the quantile plot for assessing multivariate normality of the longitudinal biomarkers. We note that except for the first part, the quantiles mostly deviate from the straight line indicating that the joint distribution of the three biomarkers deviates from a trivariate normal distribution. Therefore, the model proposed in Chapter 2 will not be appropriate in this setting.

In Figure 3.4, we show a bivariate contour plot for (log) ANC and (log) platelet count as an illustration. We show the contour for some specific density levels (i.e, 0.25, 0.5, 1, 1.5 and 2). We use solid curves for patients with an event (during treatment or in the follow-up), and the broken curves for the patients with no event during the study period. Different colors are used for different density values. We notice that except for the central part (with the higher density value) most of the contours shift up and to the right for the patients with no event. This reflects that ANC and platelet count are higher (in general) for the patients with no relapse. We cannot comment on anything similar for the central part where both the responses are observed at their median values. Along the black and the grey arrows we observe that the solid contours dominate their broken counterparts indicating that in the regions with lower values of both the responses a higher relapse rate is observed. Therefore, it is meaningful to study the effects of different quantiles of the biomarkers on the event-time instead of the effect of the mean biomarkers.

Figure 3.2, 3.3 and 3.4 jointly motivate us for a quantile-specific joint analysis of the three biomarkers and the event-time. Such an analysis gives a complete understanding of the complex association among the covariates, biomarkers and the event-time especially for our dataset where the joint distribution of the biomarkers is non-Gaussian. Quantile regression is robust than the traditional mean regression, and therefore, for the non-Gaussian setting quantile regression models are typically used for a meaningful inference (Biswas and Das, 2021 [9]).

3.3 Proposed Joint Model

In our dataset we have three biomarkers, and we define a quantile level $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$, where τ_k denotes the quantile level of the k-th biomarker, k=1,2,3. Note that our approach considers a joint quantile modeling where different biomarkers can be at different quantile levels. The quantile regression joint model (QRJM) has the following two parts: (i) a longitudinal submodel at each quantile level, and (ii) an event-time submodel which is a variant of the traditional Cox PH model. Let Y_{ijk} be the k-th

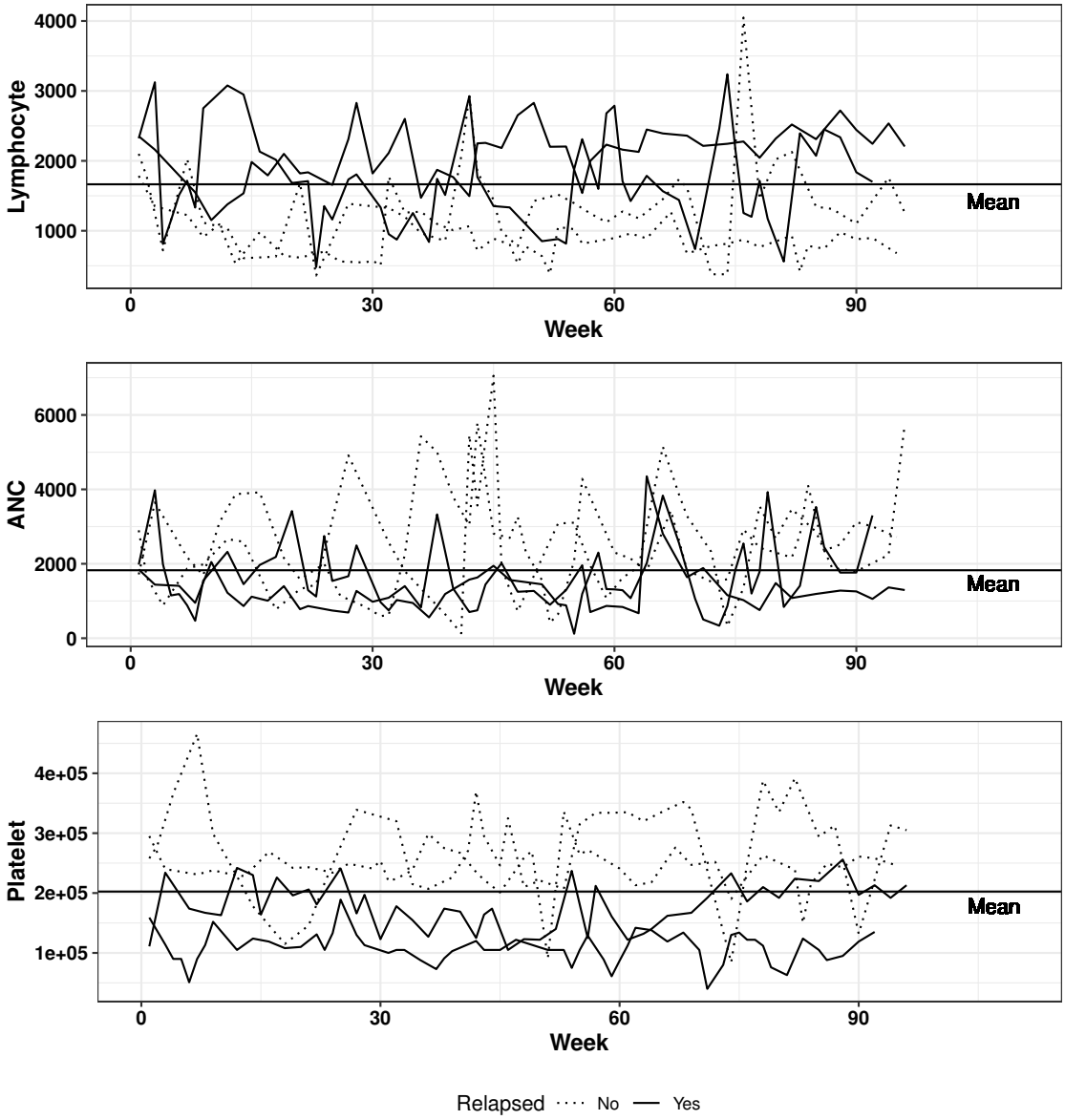


FIGURE 3.2: Longitudinal trajectories of the three biomarkers for four randomly selected children. Solid lines represent trajectories for the patients with a relapse (in the follow-up period), and the dotted lines are for those with no relapse during treatment or in the follow-up period.

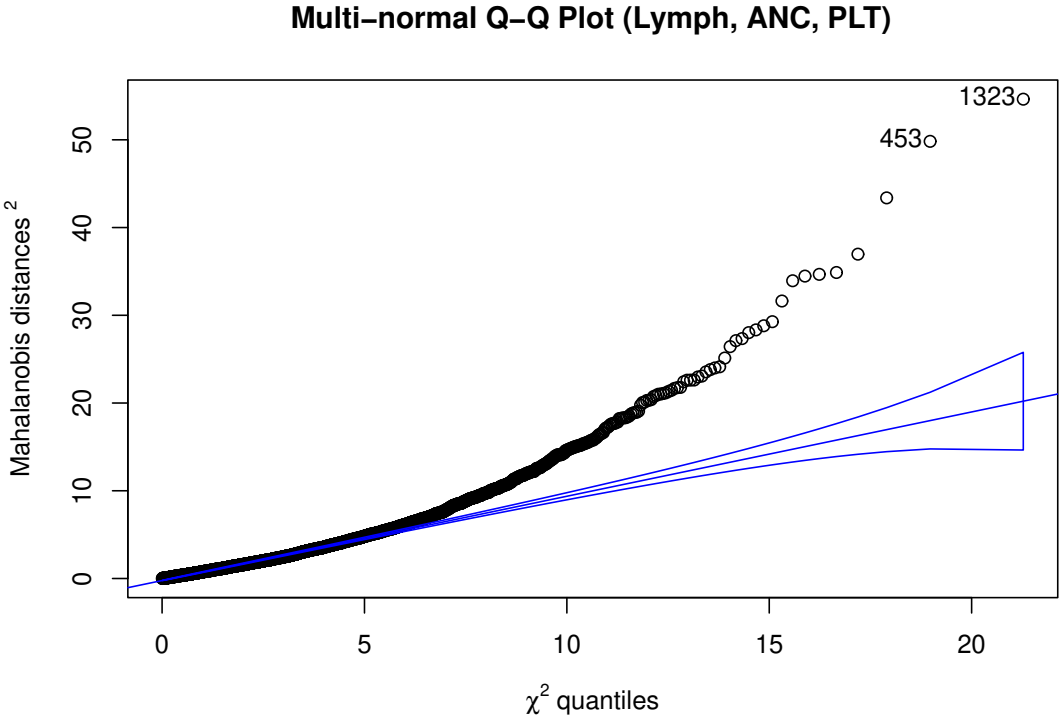


FIGURE 3.3: Quantile plot for assessing multivariate normality of the three biomarkers.

(log-transformed) biomarker ($k = 1, 2, 3$) from the i-th patient at the j-th time point, where $j = 1, 2, \ldots, t_i$; and $\lambda_i^{(\boldsymbol{\tau})}(t)$ denotes the quantile-specific hazard for the i-th individual at time t. For each individual, we have information either on the event-time (T_i) if a relapse occurs during the treatment (or in the follow-up); or on the censoring time (C_i) when the follow-up ends for the i-th patient. We define the survival time s_i=min(T_i, C_i), and also define an indicator variable δ_i=1, for $T_i < C_i$, and δ_i=0, otherwise. For each fixed quantile level $\boldsymbol{\tau}$, we jointly model the longitudinal biomarkers and the event-time as described below.

3.3.1 Longitudinal Submodel

For modeling quantiles of the longitudinal (log-transformed) biomarkers, denoted by $Q^{(\boldsymbol{\tau})}(Y_{ijk})$, we consider the following multivariate linear mixed model:

$$Q^{(\boldsymbol{\tau})}(Y_{ijk}) = f_k^{(\boldsymbol{\tau})}(t_{ij}) + \boldsymbol{\beta}_{1k}^{(\boldsymbol{\tau})T}\mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^{(\boldsymbol{\tau})T}\mathbf{z}_i + \mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij}), \tag{3.1}$$

where the general effect of time is modeled by a polynomial function (of order r) of time, i.e. $f_k^{(\boldsymbol{\tau})}(t) = \sum_{l=0}^{r} \eta_{lk}^{(\boldsymbol{\tau})} t^l$. Regression coefficients $\boldsymbol{\beta}_{1k}^{(\boldsymbol{\tau})}$ and $\boldsymbol{\beta}_{2k}^{(\boldsymbol{\tau})}$ are the quantile-specific fixed effects of time-varying covariates ($\mathbf{x}_{ij}$) and the fixed covariates ($\mathbf{z}_i$),

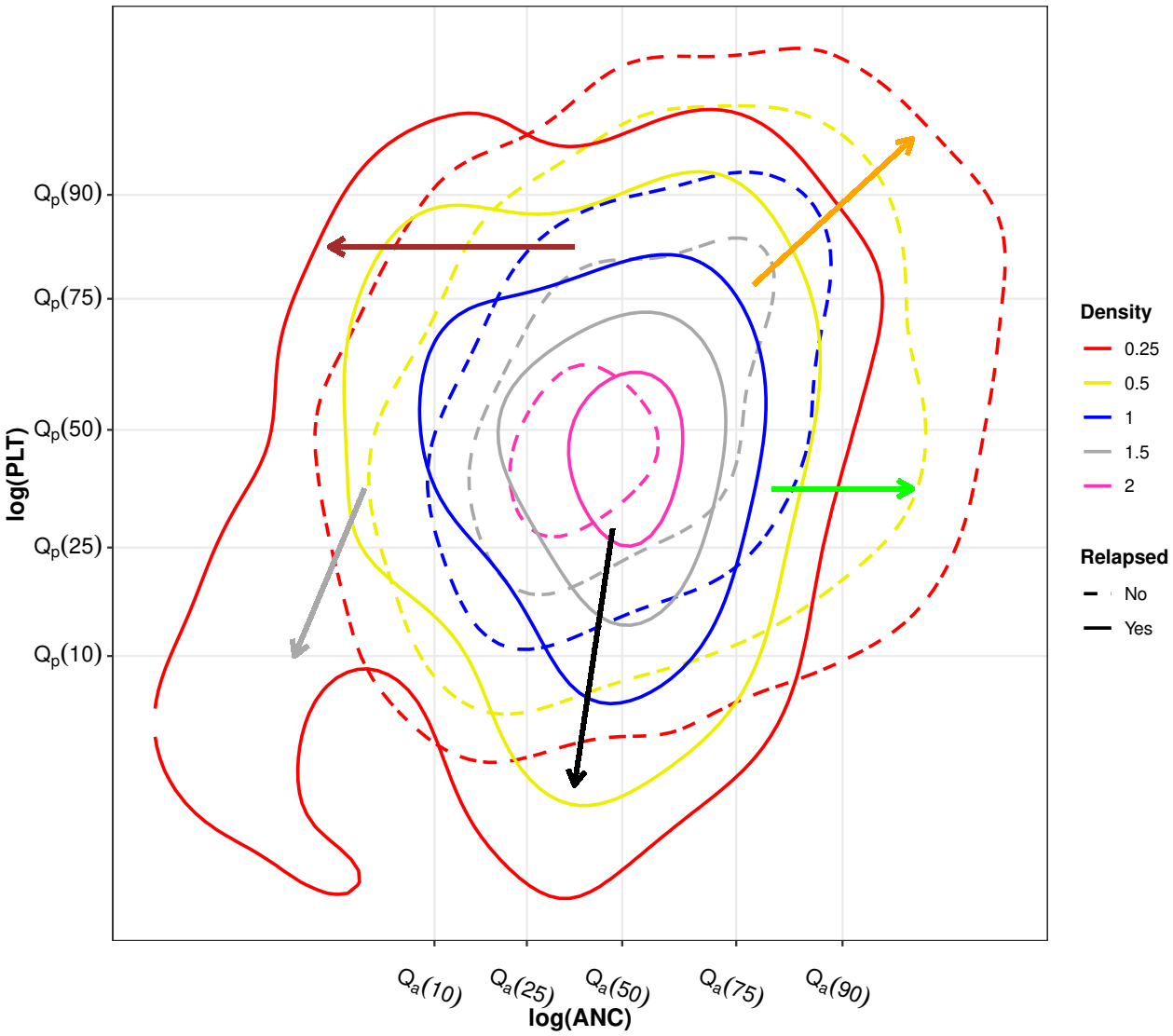


FIGURE 3.4: Contour plot of log(ANC) and log(PLT) for ALL dataset, with solid contours representing density levels for average responses for the candidates with event during their treatment/follow-up, and similarly, broken curves for the candidates with no event during the length of their study. Here, $Q_a(u)$ and $Q_p(u)$ represents the u-th quantile of (log) ANC and (log) platelets, respectively.

respectively. Subject-specific random effects, $\mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij})$, capture the longitudinal dependence among the biomarkers at different time points and also the dependence among the biomarkers over time.

We note that the model in equation (3.1) can also be expressed as follows:

$$Y_{ijk} = f_k^{(\boldsymbol{\tau})}(t_{ij}) + \boldsymbol{\beta}_{1k}^{(\boldsymbol{\tau})T}\mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^{(\boldsymbol{\tau})T}\mathbf{z}_i + \mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij}) + \epsilon_{ijk}, \tag{3.2}$$

where the random errors ϵ_{ijk} are independent observations from Asymmetric Laplace Distributions (ALD) with location parameter 0, scale parameter σ_k, and skewness parameter τ_k, for $k = 1, 2, 3$. This is similar to Geraci and Bottai (2007) [35], Kulkarni et al. (2019) [51]; and in a Bayesian setting we exploit the mixture representation of ALD as proposed in Kozumi and Kobayashi (2011) [50]. We write $\epsilon_{ijk} = \theta_{1k}e_{ijk} + \theta_{2k}\sqrt{\sigma_k e_{ijk}}v_{ijk}$, where $\theta_{1k} = \frac{1-2\tau_k}{\tau_k(1-\tau_k)}$, and $\theta_{2k} = \sqrt{\frac{2}{\tau_k(1-\tau_k)}}$, $e_{ijk} \overset{ind}{\sim} \text{Exp}(1/\sigma_k)$, and $v_{ijk} \overset{iid}{\sim} N(0,1)$, (Biswas and Das, 2021 [9]). Thus, conditional on $\mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ and e_{ijk}, the Y_{ijk} follows a normal distribution.

Subject-specific random effects $\mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ are modeled by multivariate Brownian motion which are approximated by the step functions for the computational ease. We approximate $\mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij})$ by step functions as follows: $\mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t) = \sum_{j=1}^{16} \mathbf{w}_{ijk}^{(\tau_k)}\mathbf{1}_{(t_{i,j-1}^w \le t < t_{ij}^w)}$; where $t_{i0}^w = t_{i1}$, $t_{i1}^w, \ldots t_{i,15}^w$ are 15-point Gauss-Kronrod points in the interval (t_{i1}, s_i), and $t_{i,16}^w > s_i$. Additionally, $[\mathbf{w}_{i11}^{(\boldsymbol{\tau})}, \mathbf{w}_{i12}^{(\boldsymbol{\tau})}, \mathbf{w}_{i13}^{(\boldsymbol{\tau})}]^T = \mathbf{w}_{i1}^{(\boldsymbol{\tau})} \sim N_3(\mathbf{0}, t_{i,0}^w\Sigma_{\boldsymbol{\tau}})$, and $\mathbf{w}_{ij}^{(\tau)} = \mathbf{w}_{i,j-1}^{(\tau)} + \sqrt{t_{i,j-1}^w - t_{i,j-2}^w}\boldsymbol{U}_{ij}^{(\tau)}$, where $\boldsymbol{U}_{ij}^{(\tau)} \overset{iid}{\sim} N_3(\mathbf{0}, \Sigma_{\boldsymbol{\tau}})$, $j = 2, \ldots, 16$.

An interesting feature of our model is that we do not impose any specification on the subject-specific biomarker trajectories. A simpler model with subject-specific random intercepts and random slopes (of time) could also capture the inter-biomarker dependence and the biomarker-specific longitudinal dependence (Kulkarni et al., 2019). However, that would allow the deviations of the subject-specific biomarkers from their mean to follow a straight line. But the proposed structure is quite flexible since it considers a stochastic process (multivariate Brownian motion) for the random effects. Finally, we note that for modeling the general effects of time f one can use B-splines (Devarajan and Ebrahimi, 2011 [25]; Rizopoulos, 2012 [77]) or wavelets (Moundele et al., 2019 [59]) for more flexibility. However, the raw data plots in Figure 3.1 show that a polynomial function would suffice for our dataset since the log-transformation stabilizes the variability in the biomarkers quite well.

3.3.2 Event-time Submodel

Since the relapse-time is possibly associated with the longitudinal biomarker values, we consider a joint model for the multivariate biomarkers and the event-time. However, the association (between the biomarkers and the event-time) possibly differs

from one quantile level to the other (as indicated by Figure 3.4), and hence we propose quantile-specific joint models. Yang et al. (2019) [101], Zhang and Huang (2020) [107] proposed different specifications for the quantile-specific joint modeling of longitudinal and event-time data, and we build our work on these works.

We consider a Cox PH model for quantile-specific hazards, and assume that the hazard rate for any individual at a specific time point is associated with the biomarker values at that time point, and it also depends on the time-invariant covariates. It is also assumed that the drug doses can only affect the biomarkers and not the event-time directly. For ALL dataset such assumption is valid as shown in Kundu et al. (2023) [52]. We consider the following Cox PH model for our quantile-specific joint modeling:

$$\lambda_i^{(\tau)}(t) = \lambda_0^{(\tau)}(t)\exp\left[\Psi^{(\tau)T}\boldsymbol{\mu}_i^{(\tau)}(t) + \boldsymbol{\gamma}^{(\tau)T}\mathbf{z}_i\right], \tag{3.3}$$

where $\boldsymbol{\mu}_i^{(\tau)}(t) = [\boldsymbol{\mu}_{i1}^{(\tau)}(t), \boldsymbol{\mu}_{i2}^{(\tau)}(t), \boldsymbol{\mu}_{i3}^{(\tau)}(t)]^T$, and $\boldsymbol{\mu}_{ik}^{(\tau)}(t) = f_k^{(\tau)}(t) + \boldsymbol{\beta}_{1k}^{(\tau)T}\mathbf{x}_{it} + \boldsymbol{\beta}_{2k}^{(\tau)T}\mathbf{z}_i + \mathbf{W}_{ik}^{(\tau)}(t)$; for $k = 1, 2, 3$.

Note that the biomarkers (conditional on the random effects) and the fixed covariates are considered as the predictors of the event-time at each fixed quantile level. The association parameters $\Psi^{(\tau)}$ measures the impacts of the longitudinal biomarkers on the event time, and we need to test if these parameters are statistically significant. The Baseline hazard function $\lambda_0^{(\tau)}(t)$ can be modeled in many different ways, but we follow the approach suggested in Rizopoulos (2016) [78]. We model the base-line hazard using a B-Spline, and we write $log(\lambda_0^{(\tau)}(t)) = \sum_{q=1}^{Q} \gamma_{0,q}^{(\tau)} B_q(t, \nu)$, where $B_q(t, \nu)$ is the q-th basis function of B-splines with knots $\nu_1, \nu_2, \ldots, \nu_Q$ (typically taken as equal percentiles of the event-times). Finally, $\boldsymbol{\gamma}^{(\tau)}$ measures the effects of the covariates $\mathbf{z}_i$ on the event-time.

3.3.3 Joint Likelihood and Bayesian Inference

Based on the longitudinal submodel, we get the following conditional distributions based on which the joint likelihood function can be derived.

$Y_{ijk}|e_{ijk}, \mathbf{W}_{ik}^{(\tau)}(t_{ij}) \sim N\left(f_k^{(\tau)}(t_{ij}) + \boldsymbol{\beta}_{1k}^{(\tau)T}\mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^{(\tau)T}\mathbf{z}_i + \mathbf{W}_{ik}^{(\tau)}(t_{ij}) + \theta_{1k}e_{ijk}, \theta_{2k}^2\sigma_k e_{ijk}\right)$,
$e_{ijk}|\sigma_k \sim \exp(\frac{1}{\sigma_k})$.

Let $\mathbf{W}_i^{(\tau)} = \{\mathbf{W}_{ik}^{(\tau)}\}$, $\mathbf{Y} = \{Y_{ijk}\}$, $\mathbf{s} = \{s_i\}$, and $\boldsymbol{\Theta}$ denotes the set of all model parameters (from the longitudinal and the event-time submodels). Then the joint likelihood can be written as follows:

$$L(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{s}, \mathbf{W}_i^{(\tau)}) = \prod_{i=1}^{N}\left[\prod_{j=1}^{n_i}\prod_{k=1}^{3}\left(\{f_1(Y_{ijk}|e_{ijk}, \mathbf{W}_{ik}^{(\tau)}(t_{ij})\} \times \{f_2(e_{ijk}|\sigma_k)\}\right) \times l(\mathbf{W}_i^{(\tau)}) \times l(s_i|\boldsymbol{\Theta})\right], \tag{3.4}$$

where f_1 and f_2, respectively, denote the (conditional) density of $Y_{ijk}|e_{ijk}, \mathbf{W}_{ik}^{(\boldsymbol{\tau})}(t_{ij})$; and the conditional density of $e_{ijk}|\sigma_k$. Here, $l(\mathbf{W}_i^{(\tau)}) = \frac{1}{\sqrt{2\pi|t_{i1}\boldsymbol{\Sigma}_\tau|}} \times \exp\left(-\frac{1}{2}\mathbf{w}_{i1}^T(t_{i1}\boldsymbol{\Sigma}_\tau)^{-1}\mathbf{w}_{i1}\right) \times \prod_{j=2}^{16} \frac{1}{\sqrt{2\pi|\boldsymbol{\Omega}_{ij}|}} \times \exp\left(-\frac{1}{2}(\mathbf{w}_{ij}-\mathbf{w}_{i,j-1})^T\boldsymbol{\Omega}_{ij}^{-1}(\mathbf{w}_{ij}-\mathbf{w}_{i,j-1})\right)$ is the likelihood contribution from the random effects, and $l(s_i|\boldsymbol{\Theta}) = \left(\lambda_{i\boldsymbol{\tau}}(s_i)\right)^{\delta_i} \times \exp\left(-\int_0^{s_i}\lambda_{i\boldsymbol{\tau}}(t)dt\right)$ is the likelihood contribution (for the i-th individual) from the event-time submodel. Note that, Ω_{ij} is the 3×3 variance-covariance matrix for dependent Weiner process, i.e $\mathbf{w}_{ij}$, where $\Omega_{ij} = (t_{ij}^w - t_{i,j-1}^w)\boldsymbol{\Sigma}_\tau$; $\ j = 1, \ldots, 15$.

We consider a Bayesian approach where some prior distributions are assumed for $\boldsymbol{\Theta}$, and then we consider the joint posterior distribution $\pi(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{s}) \propto L(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{s}, \mathbf{W}_i^{(\boldsymbol{\tau})}) \times \pi(\boldsymbol{\Theta})$. Assuming independent prior distributions for different model parameters we sample from the joint posterior distribution using Markov Chain Monte Carlo (MCMC) algorithm, and the model parameters are estimated by their respective sample means. All our computations are done in *R* using *JAGS 4.3.0*.

3.4 ALL Data Analysis

3.4.1 Prior Specification and Computational Details

We use a Bayesian approach for our computation and data analysis. We specify prior distributions for the model parameters based on the existing literature (mostly the flat priors). For each component of the vector $\boldsymbol{\beta}_{1k}^{(\boldsymbol{\tau})}$ and $\boldsymbol{\beta}_{2k}^{(\boldsymbol{\tau})T}$ we specify a $N(0, 1000)$ prior. We specify the same prior (i.e. a $N(0, 1000)$) for $\eta_{lk}^{(\boldsymbol{\tau})}$, $l = 0, 1, \ldots, r$; and for each component of the vector $\Psi^{(\boldsymbol{\tau})}$ we also specify a $N(0, 1000)$ prior. The same prior is used for $\boldsymbol{\gamma}^{(\boldsymbol{\tau})}$. For the inverse of the covariance matrix Σ_τ we specify a Wishart(I_4,3) prior, and for σ_k we specify an Inverse Gamma (0.01,0.01) prior. For modeling the baseline hazard function $\lambda_0^{(\boldsymbol{\tau})}(t)$ we use cubic B-splines as mentioned in Section 3.2. However, for selecting the optimal number of knots we use the recommendation given in Rizopoulos (2016). We use a large number of knots, and then specify (shrinkage) priors for the (B spline) coefficients $\gamma_{0,q}^{(\tau)}$. For our analysis, we consider 15 knots and then specify a Laplace (0, κ) prior, and for κ we consider a Gamma (0.5,0.5) prior.

For estimating the model parameters, we implement MCMC algorithm. We run 20,000 MCMC iterations for each of the 5 independent chains, and the first 10,000 iterations are discarded as "burn-in" in each chain. Then we thin the chains by saving every 5-th iteration which results in a total of 10,000 iterations from all 5 chains. Model parameters are estimated by their respective sample means based on 10,000 MCMC iterations. Convergence of the Markov Chains are assessed by computing scale reduction factors (Brooks and Gelman, 1998 [12]). For our computation, scale reduction factors for all the model parameters are smaller than 1.1 (which indicates a good convergence). Trace plots for some of the model parameters are given in Figure

3.5. These trace plots and the computed scale reduction factors indicate that the chains converge well. Similar results are obtained for the other model parameters as well (results not shown).

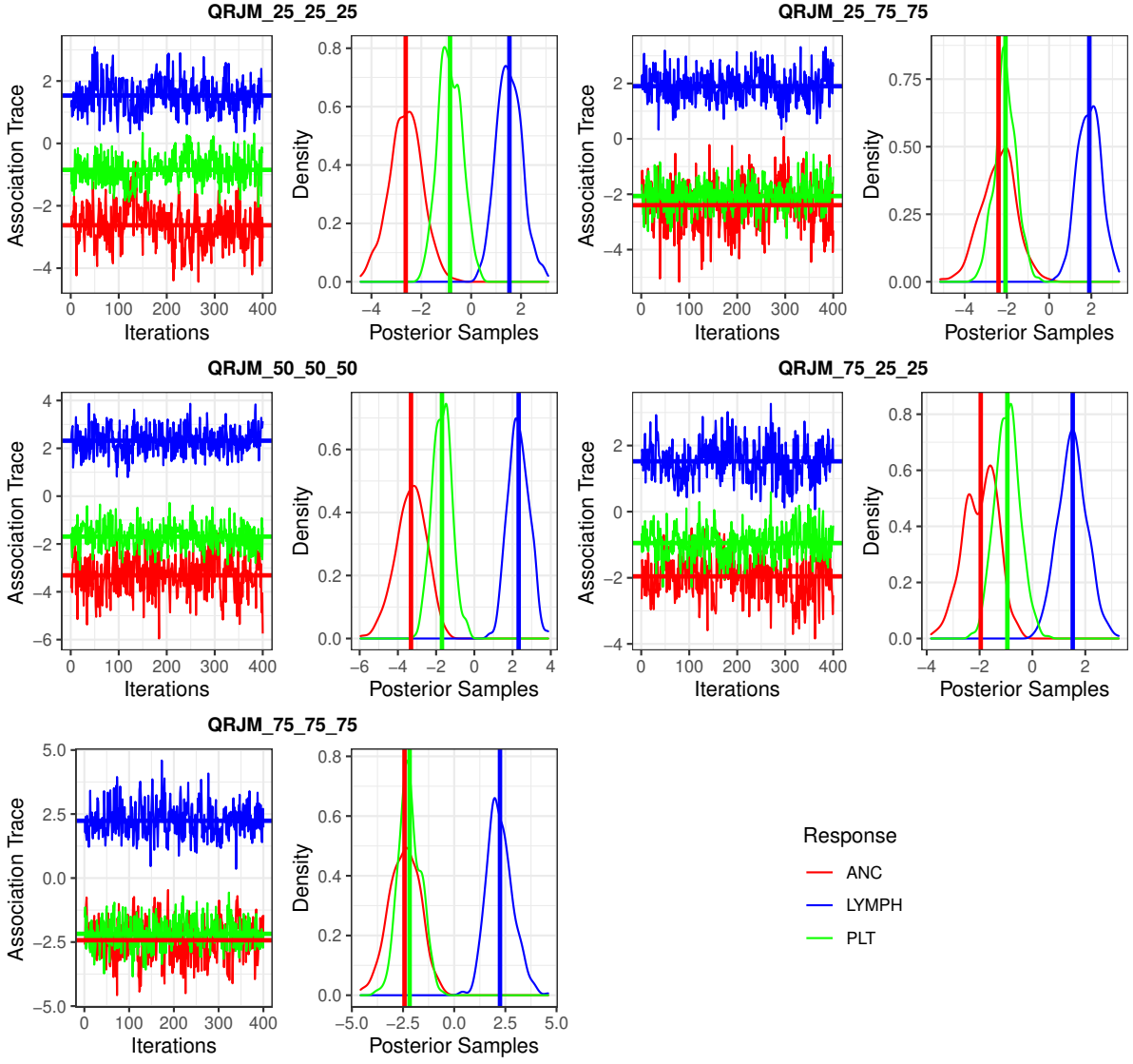


FIGURE 3.5: Estimated posterior density and trace plots for the three association parameters in ALL data analysis.

In this analysis, we consider five different choices for $\boldsymbol{\tau}$, i.e. $\boldsymbol{\tau} = (25, 25, 25)$, $(25, 75, 75)$, $(50, 50, 50)$, $(75, 25, 25)$, $(75, 75, 75)$. Note that $\boldsymbol{\tau}$=(25,25,25), (75,75,75), and (50,50,50) consider the cases where all three biomarkers are at lower levels, upper levels, and at median levels, respectively. Two extreme cases are also considered, i.e. $\boldsymbol{\tau}$=(25,75,75) and (75,25,25), which represent a lower level of lymphocyte count with upper levels of ANC and platelet count, and a upper level of lymphocyte count with lower levels of ANC and platelet count. We avoid the extreme quantiles, i.e. (10,10,10) or (90,90,90) since they result in the inconsistent estimates.

To determine the optimal order r of the polynomial function $f_k^{(\boldsymbol{\tau})}(t_{ij})$ (in equation

(3.2)) we consider a linear, a quadratic and a cubic function, and fit the model for the five quantile levels. We compute three standard measures for model selection, i.e. BIC, DIC and LPML. Table 3.2 summarizes the results where we report the average values of different measures (averaged over the quantile levels). It is noted that the smallest value of BIC and the largest value of LPML are obtained for r=2, and the DIC value for r=1 is slightly smaller than that of r=2. Hence, we select r=2, and perform our analysis accordingly.

TABLE 3.2: Average BIC, DIC and LPML values for selecting the optimal order (r) in ALL data analysis.

r	BIC	DIC	LPML
1	5821.39	27.18	-203.43
2	5427.15	28.36	-180.51
3	6328.46	39.72	-217.66

TABLE 3.3: BIC and DIC values for the proposed joint modeling and separate modeling for five different quantile levels.

Quantile level	Joint modeling		Separate modeling	
$\boldsymbol{\tau}$	BIC	DIC	BIC	DIC
(25,25,25)	5318.26	32.57	5816.29	39.72
(50,50,50)	5519.04	35.10	6115.41	46.13
(75,75,75)	5927.82	29.22	6327.33	37.44
(25,75,75)	5815.23	33.45	6514.17	52.38
(75,25,25)	6123.51	41.19	6552.04	55.19

3.4.2 Results

3.4.2.1 Model Comparison

We assess the effectiveness of the proposed joint modeling by comparing it with separate modeling of the longitudinal biomarkers and event-time. While considering separate modeling, we first model three biomarkers jointly using the linear mixed model given in equation (3.2) for each quantile level $\boldsymbol{\tau}$. Model parameters are estimated based on the joint posterior distribution. Then, the PH model given in equation (3.3) is used for modeling the event-time where $\boldsymbol{\mu}_i^{(\boldsymbol{\tau})}(t)$ is replaced with its estimated value $\widehat{\boldsymbol{\mu}}_i^{(\boldsymbol{\tau})}(t) = \hat{f}_k^{(\boldsymbol{\tau})}(t) + \hat{\boldsymbol{\beta}}_{1k}^{(\boldsymbol{\tau})T}\mathbf{x}_{it} + \hat{\boldsymbol{\beta}}_{2k}^{(\boldsymbol{\tau})T}\mathbf{z}_i + \widehat{\mathbf{W}}_{ik}^{(\boldsymbol{\tau})}(t)$.

For each of the five quantile levels (as discussed in Section 3.4.1), we fit two separate models, and the proposed joint model. We compute the BIC and the DIC values for the joint and the separate models. DIC for our setting is similar to Das and Daniels (2014) [20], and we use their approach for computing it for random effects models. Note that for separate modeling the overall BIC (and DIC) values are obtained by adding the BIC (and DIC) values from the corresponding longitudinal

models and the event-time model. In Table 3.3 we report the BIC and the DIC values for the joint and separate modeling for the five different quantile levels. We note that across all quantiles we obtain smaller BIC and DIC values for the proposed joint modeling. This indicates that the proposed joint model fits our data better than a separate modeling.

3.4.2.2 Effects of different covariates

In Figure 3.6, we show the significant fixed covariates at different quantile levels for three biomarkers and the event-time. Not that the significance of a covariate is assessed by the estimated 95% credible interval of the corresponding regression coefficients (whether it contains a zero or not). We see that age is a significant predictor (with positive effects) for the lymphocyte count at all the five quantile levels, but it is significant (with a negative effect) for the platelet count only at the level $\tau =$ (75,25,25). Bulky disease is significant (with negative effects) at all the five levels for ANC and platelet counts, but for the lymphocyte count it is significant (with a negative effect) only at the median level. While gender is mostly significant (at four quantile levels) with positive effects for the lymphocyte count and the event-time, for ANC it is significant (positively) only at one level $\tau =$(75,25,25), and at two levels for the platelet counts. The interesting thing to note here is, no predictor is selected as significant for all the three biomarkers (and the event-time) at all the five quantile levels, and also there is no predictor which is not-significant (for all outcomes) at all the five quantile levels. This illustrates that the set of covariates affecting the biomarkers change from one quantile level to the other. This finding justifies the necessity for a quantile-specific inference for the dataset under consideration.

In Figure 3.7, we show the estimated effects of the two drugs on three biomarkers and also their 95% Bayesian credible intervals. We note that effects of 6MP are negative for the lymphocyte counts and the credible intervals do not contain a zero across all quantile levels. This indicates that 6MP is indeed quite effective in reducing the lymphocyte counts. Effects of MTx are close to zero for the lymphocyte counts, and the credible intervals contain zero indicating that MTx does not help in reducing the lymphocyte count across all quantiles. For the platelet count we note that the effects of 6MP are all positive and the credible intervals do not contain zero at $\tau =$(25,75,75), and $\tau =$(75,75,75). However, for MTx the corresponding credible intervals contain zeros which again indicates that MTx does not affect the platelet count. On the other hand, for ANC the estimated effects of MTx are all positive and the credible intervals do not contain zeros. But the credible intervals for the effects of 6MP on ANC mostly contain zeros (except at the level $\tau =$(25,25,25)). This indicates that MTx is quite effective in increasing the neutrophil counts, but 6MP is not effective for that. To summarize, 6MP is not only effective in reducing the Lymphocyte count irrespective of the levels of other two biomarkers, but is also responsible for further improving the platelet counts for patients who have platelets counts relatively on a higher side.

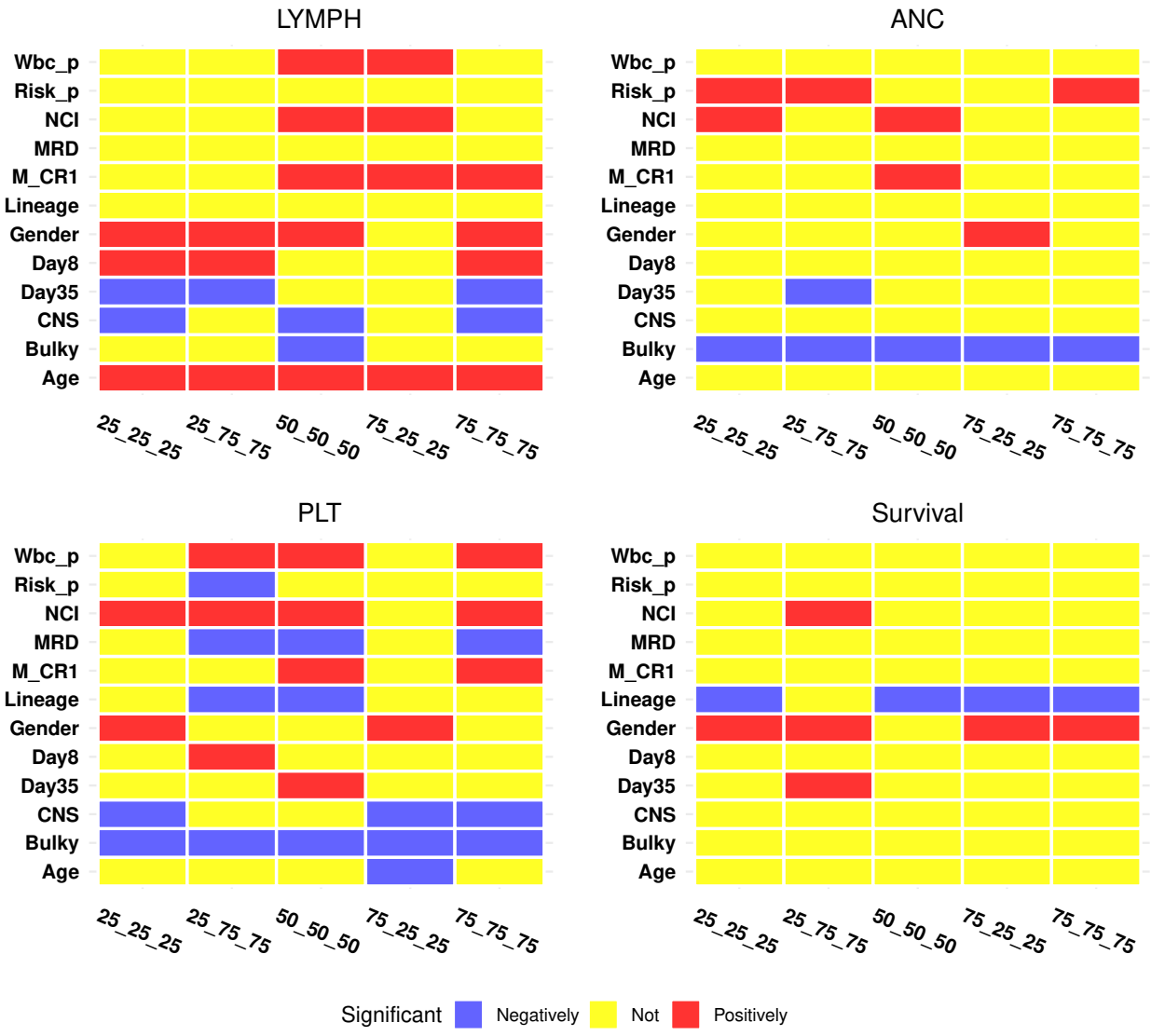


FIGURE 3.6: Quantile-specific significance of the fixed covariates for different submodels

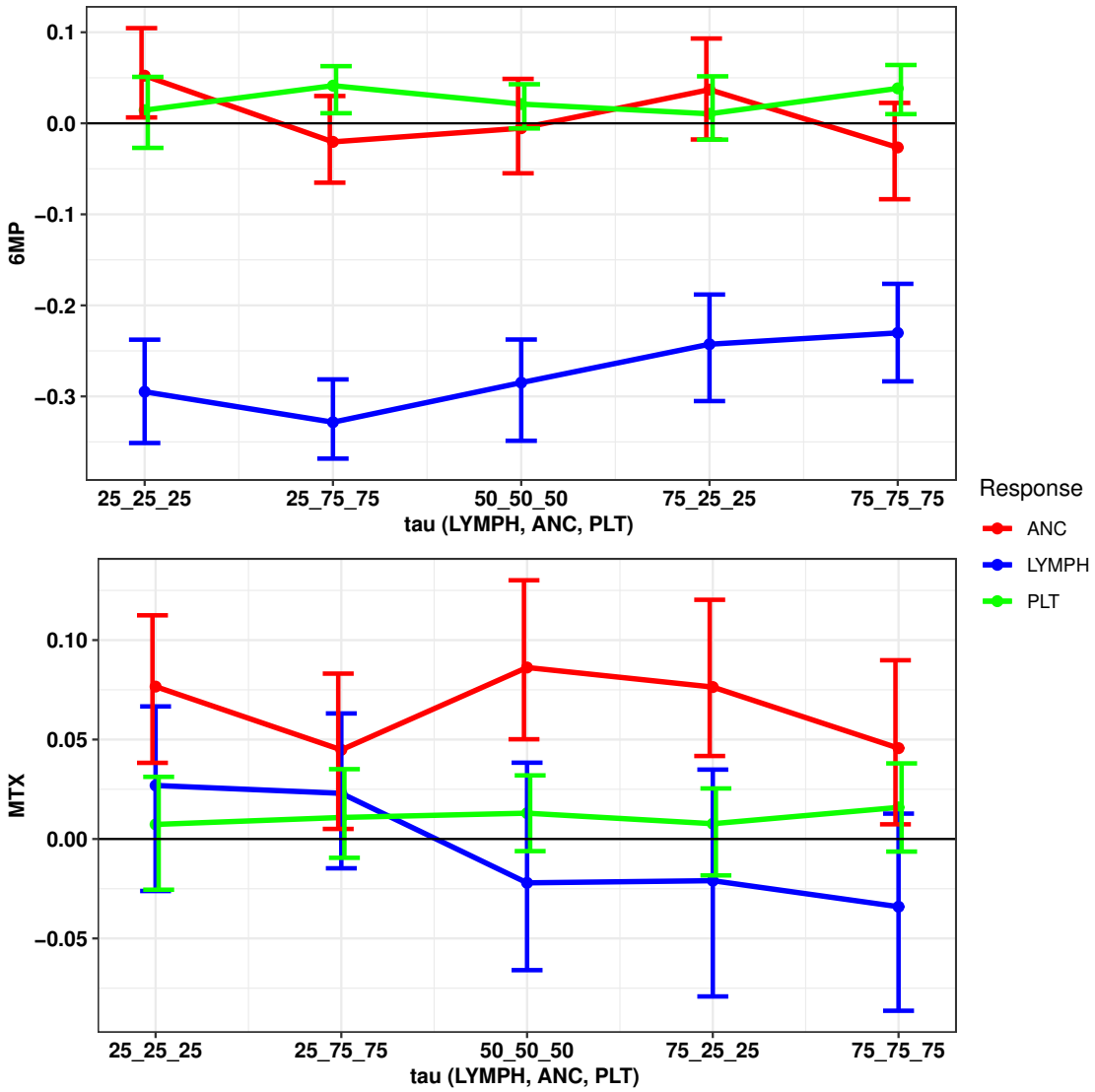


FIGURE 3.7: Estimate and 95% credible interval for the quantile-specific effects of the drugs on the three biomarkers.

MTx on the other hand plays an important role in increasing ANC across all quantile combinations. Based on these results we conclude that the doses of 6MP and MTx should be recommended based on the levels of the biomarkers at any time point.

In Tables 3.4-3.13, we report the estimated covariate effects, the respective Monte Carlo Standard Errors (MCSE), and 95% estimated Bayesian credible intervals for the three biomarkers at the five different quantile levels.

TABLE 3.4: Estimated covariate effects (with the monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the three biomarkers at τ =(25,25,25).

	Lymphocyte count		Neutrophil count		Platelet count	
Covariate	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI
6MP dose	-0.295(0.026)	(-0.351,-0.238)	0.052(0.026)	(0.006,0.105)	0.014(0.021)	(-0.027,0.051)
MTx dose	0.027(0.022)	(-0.026,0.067)	0.077(0.019)	(0.038,0.112)	0.007(0.014)	(-0.026,0.031)
Age at diagnosis	0.104(0.036)	(0.043,0.178)	0.078(0.05)	(-0.037,0.152)	0.017(0.056)	(-0.085,0.088)
WBC at presentation	0.04(0.032)	(-0.012,0.101)	0.041(0.05)	(-0.03,0.134)	0.058(0.046)	(-0.023,0.12)
Gender	0.135(0.066)	(0.038,0.272)	0.061(0.052)	(-0.035,0.17)	0.109(0.034)	(0.037,0.167)
Lineage	-0.075(0.064)	(-0.193,0.056)	-0.047(0.078)	(-0.173,0.134)	-0.053(0.062)	(-0.176,0.031)
NCI risk group	0.042(0.035)	(-0.024,0.106)	0.176(0.091)	(0.043,0.329)	0.238(0.087)	(0.124,0.365)
Bulky disease	-0.097(0.068)	(-0.212,0.017)	-0.246(0.04)	(-0.323,-0.175)	-0.166(0.024)	(-0.203,-0.105)
CNS disease	-0.087(0.061)	(-0.181,-0.009)	-0.008(0.109)	(-0.168,0.151)	-0.096(0.047)	(-0.18,-0.019)
Risk at presentation	-0.008(0.047)	(-0.08,0.08)	0.092(0.047)	(0.023,0.208)	0.017(0.016)	(-0.007,0.056)
Day 8 risk	0.119(0.063)	(0.024,0.253)	0.01(0.051)	(-0.077,0.097)	0.054(0.035)	(-0.007,0.112)
Day 35 risk	-0.043(0.03)	(-0.099,-0.002)	-0.003(0.044)	(-0.085,0.068)	0.028(0.036)	(-0.031,0.092)
Morphological remission	0.08(0.065)	(-0.012,0.216)	0.075(0.093)	(-0.056,0.254)	0.035(0.03)	(-0.016,0.088)
MRD status	0.06(0.046)	(-0.017,0.126)	0.078(0.051)	(-0.014,0.163)	-0.033(0.042)	(-0.105,0.03)

TABLE 3.5: Estimated covariate effects (with monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the event-time at τ =(25,25,25).

Covariate	Est.(MCSE)	95% CI
Lymphocyte count	1.534(0.506)	(0.641,2.574)
Neutrophil count	-2.625(0.654)	(-3.954,-1.325)
Platelet count	-0.847(0.46)	(-1.706,0.024)
Age at diagnosis	0.23(0.221)	(-0.192,0.663)
WBC at presentation	0.102(0.163)	(-0.206,0.405)
Gender	0.752(0.334)	(0.158,1.442)
Lineage	-1.848(0.566)	(-2.999,-0.767)
NCI risk group	0.384(0.385)	(-0.393,1.13)
Bulky disease	-0.162(0.267)	(-0.713,0.365)
CNS disease	0.116(0.295)	(-0.454,0.717)
Risk at presentation	0.107(0.265)	(-0.389,0.626)
Day 8 risk	-0.118(0.354)	(-0.797,0.556)
Day 35 risk	0.152(0.169)	(-0.172,0.492)
Morphological remission	0.271(0.581)	(-0.863,1.408)
MRD status	-0.282(0.295)	(-0.829,0.313)

TABLE 3.6: Estimated covariate effects (with the monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the three biomarkers at the median level, i.e. at τ=(50,50,50).

	Lymphocyte count		Neutrophil count		Platelet count	
Covariate	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI
6MP dose	-0.285(0.028)	(-0.349,-0.238)	-0.005(0.027)	(-0.055,0.049)	0.021(0.013)	(-0.006,0.043)
MTx dose	-0.022(0.027)	(-0.066,0.038)	0.086(0.021)	(0.05,0.13)	0.013(0.01)	(-0.006,0.032)
Age at diagnosis	0.115(0.034)	(0.061,0.178)	0.036(0.037)	(-0.033,0.103)	-0.016(0.019)	(-0.06,0.019)
WBC at presentation	0.047(0.015)	(0.021,0.087)	0.026(0.016)	(-0.004,0.056)	0.058(0.014)	(0.03,0.081)
Gender	0.086(0.046)	(0,0.172)	-0.018(0.039)	(-0.093,0.053)	0.067(0.046)	(-0.009,0.146)
Lineage	-0.065(0.045)	(-0.134,0.034)	-0.032(0.04)	(-0.118,0.044)	-0.071(0.03)	(-0.129,-0.02)
NCI risk group	0.067(0.034)	(0.004,0.117)	0.077(0.039)	(0.022,0.165)	0.233(0.023)	((0.194,0.271)
Bulky disease	-0.109(0.027)	(-0.149,-0.05)	-0.173(0.04)	(-0.262,-0.106)	-0.109(0.038)	(-0.171,-0.048)
CNS disease	-0.1(0.047)	(-0.191,-0.034)	0.056(0.047)	(-0.028,0.138)	-0.039(0.039)	(-0.112,0.018)
Risk at presentation	-0.015(0.042)	(-0.12,0.053)	0.066(0.054)	(-0.015,0.151)	-0.031(0.023)	(-0.074,0.01)
Day 8 risk	0.049(0.047)	(-0.057,0.113)	-0.03(0.082)	(-0.146,0.132)	0.021(0.032)	(-0.03,0.072)
Day 35 risk	-0.03(0.017)	(-0.06,0.001)	-0.002(0.035)	(-0.074,0.054)	0.04(0.02)	(0.006,0.073)
Morphological remission	0.164(0.106)	(0.003,0.391)	0.138(0.047)	(0.066,0.251)	0.096(0.035)	(0.025,0.153)
MRD status	0.001(0.049)	(-0.093,0.09)	0.001(0.029)	(-0.05,0.059)	-0.057(0.017)	(-0.088,-0.019)

TABLE 3.7: Estimated covariate effects (with monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the event-time at the median level, i.e. at τ=(50,50,50).

Covariate	Est.(MCSE)	95% CI
Lymphocyte count	2.318(0.542)	(1.353,3.286)
Neutrophil count	-3.313(0.789)	(-4.973,-1.9)
Platelet count	-1.693(0.499)	(-2.655,-0.691)
Age at diagnosis	0.189(0.218)	(-0.231,0.62)
WBC at presentation	0.041(0.156)	(-0.292,0.34)
Gender	0.625(0.355)	(-0.044,1.308)
Lineage	-1.507(0.554)	(-2.722,-0.559)
NCI risk group	0.536(0.371)	(-0.183,1.246)
Bulky disease	-0.231(0.292)	(-0.768,0.384)
CNS disease	0.235(0.309)	(-0.33,0.876)
Risk at presentation	0.262(0.263)	(-0.228,0.773)
Day 8 risk	-0.344(0.368)	(-1.022,0.354)
Day 35 risk	0.304(0.185)	(-0.037,0.668)
Morphological remission	0.304(0.647)	(-0.822,1.487)
MRD status	-0.312(0.289)	(-0.889,0.236)

TABLE 3.8: Estimated covariate effects (with the monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the three biomarkers at τ =(25,75,75).

	Lymphocyte count		Neutrophil count		Platelet count	
Covariate	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI
6MP dose	-0.328(0.024)	(-0.368,-0.281)	-0.02(0.025)	(-0.065,0.03)	0.041(0.013)	(0.011,0.063)
MTx dose	0.023(0.021)	(-0.015,0.063)	0.045(0.019)	(0.005,0.083)	0.011(0.012)	(-0.009,0.035)
Age at diagnosis	0.133(0.03)	(0.084,0.194)	0.041(0.032)	(-0.017,0.107)	-0.022(0.033)	(-0.08,0.033)
WBC at presentation	0.038(0.025)	(-0.014,0.076)	0.01(0.017)	(-0.019,0.046)	0.046(0.01)	(0.025,0.066)
Gender	0.137(0.059)	(0.005,0.22)	-0.02(0.038)	(-0.097,0.053)	0.019(0.033)	(-0.06,0.086)
Lineage	-0.088(0.08)	(-0.207,0.034)	0(0.062)	(-0.104,0.111)	-0.081(0.027)	(-0.125,-0.037)
NCI risk group	0.102(0.076)	(-0.053,0.241)	0.072(0.071)	(-0.034,0.199)	0.179(0.041)	(0.097,0.228)
Bulky disease	-0.066(0.054)	(-0.161,0.006)	-0.171(0.041)	(-0.267,-0.105)	-0.049(0.026)	(-0.115,-0.009)
CNS disease	-0.015(0.062)	(-0.124,0.054)	0.055(0.038)	(-0.018,0.116)	-0.044(0.031)	(-0.098,0.015)
Risk at presentation	0.028(0.035)	(-0.047,0.082)	0.086(0.025)	(0.033,0.132)	-0.033(0.019)	(-0.064,-0.002)
Day 8 risk	0.11(0.036)	(0.055,0.193)	0.036(0.024)	(-0.011,0.083)	0.071(0.048)	(0.005,0.165)
Day 35 risk	-0.057(0.025)	(-0.123,-0.023)	-0.054(0.024)	(-0.113,-0.016)	-0.007(0.023)	(-0.045,0.032)
Morphological remission	-0.026(0.057)	(-0.135,0.079)	0.01(0.095)	(-0.165,0.174)	-0.011(0.055)	(-0.121,0.092)
MRD status	-0.019(0.02)	(-0.058,0.017)	-0.041(0.029)	(-0.097,0.016)	-0.049(0.02)	(-0.101,-0.021)

TABLE 3.9: Estimated covariate effects (with monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the event-time at τ =(25,75,75).

Covariate	Est.(MCSE)	95% CI
Lymphocyte count	1.901(0.555)	(0.781,3.019)
Neutrophil count	-2.396(0.825)	(-4.075,-0.886)
Platelet count	-2.066(0.509)	(-3.052,-1.046)
Age at diagnosis	0.238(0.228)	(-0.208,0.654)
WBC at presentation	0.069(0.155)	(-0.233,0.351)
Gender	0.743(0.325)	(0.144,1.422)
Lineage	-1.052(0.519)	(-1.97,0.029)
NCI risk group	0.968(0.383)	(0.257,1.813)
Bulky disease	0.206(0.284)	(-0.385,0.759)
CNS disease	0.075(0.295)	(-0.49,0.686)
Risk at presentation	0.229(0.281)	(-0.308,0.753)
Day 8 risk	-0.402(0.375)	(-1.177,0.341)
Day 35 risk	0.48(0.205)	(0.074,0.849)
Morphological remission	0.316(0.711)	(-1.106,1.736)
MRD status	-0.504(0.284)	(-1.052,0.11)

TABLE 3.10: Estimated covariate effects (with the monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the three biomarkers at τ =(75,25,25).

	Lymphocyte count		Neutrophil count		Platelet count	
Covariate	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI
6MP dose	-0.243(0.029)	(-0.305,-0.188)	0.037(0.028)	(-0.018,0.093)	0.01(0.018)	(-0.018,0.052)
MTx dose	-0.021(0.028)	(-0.079,0.035)	0.076(0.02)	(0.042,0.12)	0.008(0.011)	(-0.018,0.025)
Age at diagnosis	0.126(0.027)	(0.067,0.172)	-0.019(0.045)	(-0.095,0.071)	-0.09(0.031)	(-0.151,-0.025)
WBC at presentation	0.063(0.021)	(0.024,0.106)	-0.027(0.025)	(-0.068,0.027)	-0.019(0.018)	(-0.05,0.014)
Gender	0.087(0.069)	(-0.023,0.225)	0.1(0.048)	(0.015,0.19)	0.156(0.042)	(0.084,0.244)
Lineage	-0.046(0.049)	(-0.155,0.029)	-0.023(0.061)	(-0.142,0.088)	0.04(0.083)	(-0.084,0.171)
NCI risk group	0.164(0.056)	(0.104,0.288)	-0.022(0.097)	(-0.194,0.124)	0.017(0.049)	(-0.071,0.097)
Bulky disease	-0.056(0.036)	(-0.107,0.011)	-0.22(0.043)	(-0.302,-0.152)	-0.09(0.045)	(-0.167,-0.016)
CNS disease	-0.035(0.036)	(-0.105,0.027)	0.065(0.038)	(-0.012,0.127)	-0.067(0.041)	(-0.146,-0.006)
Risk at presentation	0(0.038)	(-0.066,0.063)	0.084(0.045)	(-0.008,0.165)	0.055(0.032)	(-0.008,0.101)
Day 8 risk	0.053(0.074)	(-0.05,0.173)	-0.043(0.055)	(-0.133,0.043)	-0.017(0.051)	(-0.1,0.059)
Day 35 risk	-0.025(0.018)	(-0.069,0.01)	-0.024(0.032)	(-0.086,0.028)	0.012(0.022)	(-0.023,0.053)
Morphological remission	0.21(0.055)	(0.096,0.305)	0.146(0.154)	(-0.105,0.413)	0.055(0.064)	(-0.034,0.176)
MRD status	-0.023(0.036)	(-0.089,0.046)	0.035(0.04)	(-0.044,0.12)	-0.047(0.025)	(-0.089,0.006)

TABLE 3.11: Estimated covariate effects (with monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the event-time at τ =(75,25,25).

Covariate	Est.(MCSE)	95% CI
Lymphocyte count	1.524(0.546)	(0.475,2.583)
Neutrophil count	-1.957(0.63)	(-3.219,-0.802)
Platelet count	-0.949(0.466)	(-1.809,-0.013)
Age at diagnosis	0.137(0.238)	(-0.329,0.585)
WBC at presentation	0.018(0.165)	(-0.292,0.35)
Gender	0.791(0.349)	(0.149,1.454)
Lineage	-1.685(0.578)	(-2.753,-0.63)
NCI risk group	0.172(0.42)	(-0.626,0.958)
Bulky disease	-0.231(0.262)	(-0.762,0.271)
CNS disease	0.097(0.295)	(-0.463,0.69)
Risk at presentation	0.125(0.233)	(-0.347,0.555)
Day 8 risk	-0.035(0.349)	(-0.769,0.654)
Day 35 risk	0.072(0.194)	(-0.261,0.41)
Morphological remission	0.131(0.609)	(-1.005,1.449)
MRD status	-0.228(0.274)	(-0.746,0.35)

TABLE 3.12: Estimated covariate effects (with the monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the three biomarkers at $\boldsymbol{\tau}$ =(75,75,75).

	Lymphocyte count		Neutrophil count		Platelet count	
Covariate	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI	Est.(MCSE)	95% CI
6MP dose	-0.23(0.026)	(-0.283,-0.176)	-0.027(0.027)	(-0.083,0.022)	0.038(0.015)	(0.01,0.064)
MTx dose	-0.034(0.026)	(-0.086,0.013)	0.046(0.022)	(0.007,0.09)	0.016(0.012)	(-0.006,0.038)
Age at diagnosis	0.117(0.058)	(0.037,0.228)	0.035(0.045)	(-0.04,0.132)	-0.023(0.041)	(-0.09,0.044)
WBC at presentation	0.047(0.022)	(-0.005,0.081)	0.004(0.018)	(-0.038,0.039)	0.035(0.016)	(0.003,0.064)
Gender	0.105(0.045)	(0.009,0.179)	-0.017(0.051)	(-0.109,0.1)	0.035(0.029)	(-0.024,0.094)
Lineage	-0.03(0.123)	(-0.236,0.181)	-0.021(0.06)	(-0.16,0.065)	-0.045(0.05)	(-0.119,0.033)
NCI risk group	0.117(0.061)	(-0.012,0.21)	0.048(0.071)	(-0.055,0.173)	0.185(0.067)	(0.089,0.267)
Bulky disease	-0.05(0.028)	(-0.104,0)	-0.17(0.034)	(-0.228,-0.099)	-0.059(0.025)	(-0.097,-0.008)
CNS disease	-0.173(0.049)	(-0.264,-0.096)	-0.028(0.037)	(-0.101,0.043)	-0.088(0.02)	(-0.12,-0.055)
Risk at presentation	-0.005(0.044)	(-0.074,0.069)	0.077(0.039)	(0.007,0.149)	-0.033(0.019)	(-0.061,0.004)
Day 8 risk	0.082(0.032)	(0.044,0.154)	0.001(0.072)	(-0.132,0.109)	0.037(0.052)	(-0.04,0.111)
Day 35 risk	-0.055(0.021)	(-0.104,-0.021)	-0.043(0.035)	(-0.105,0.013)	0.011(0.021)	(-0.022,0.062)
Morphological remission	0.168(0.053)	(0.078,0.267)	0.085(0.061)	(-0.023,0.237)	0.103(0.043)	(0.012,0.166)
MRD status	-0.002(0.023)	(-0.044,0.04)	-0.033(0.04)	(-0.121,0.033)	-0.084(0.028)	(-0.131,-0.037)

TABLE 3.13: Estimated covariate effects (with monte carlo standard errors (MCSE) and 95% Bayesian credible intervals) on the event-time at $\boldsymbol{\tau}$ =(75,75,75).

Covariate	Est.(MCSE)	95% CI
Lymphocyte count	2.237(0.631)	(1.171,3.576)
Neutrophil count	-2.428(0.748)	(-3.894,-1.009)
Platelet count	-2.178(0.543)	(-3.158,-1.115)
Age at diagnosis	0.175(0.226)	(-0.262,0.63)
WBC at presentation	0.007(0.162)	(-0.316,0.316)
Gender	0.702(0.319)	(0.083,1.323)
Lineage	-1.159(0.578)	(-2.409,-0.042)
NCI risk group	0.697(0.378)	(0,1.459)
Bulky disease	0.018(0.281)	(-0.495,0.568)
CNS disease	0.238(0.312)	(-0.37,0.867)
Risk at presentation	0.3(0.288)	(-0.298,0.854)
Day 8 risk	-0.331(0.387)	(-1.089,0.414)
Day 35 risk	0.35(0.191)	(-0.015,0.722)
Morphological remission	0.056(0.692)	(-1.229,1.343)
MRD status	-0.445(0.293)	(-1.018,0.136)

3.4.2.3 Association parameters and non-relapse probabilities

In Figure 3.8, we show the estimated association parameters (measuring the effects of the biomarkers on the hazard, the risk of an instantaneous relapse) at different quantile levels. We note that at all five quantile levels the estimated association parameter for the lymphocyte count is positive and their respective credible intervals

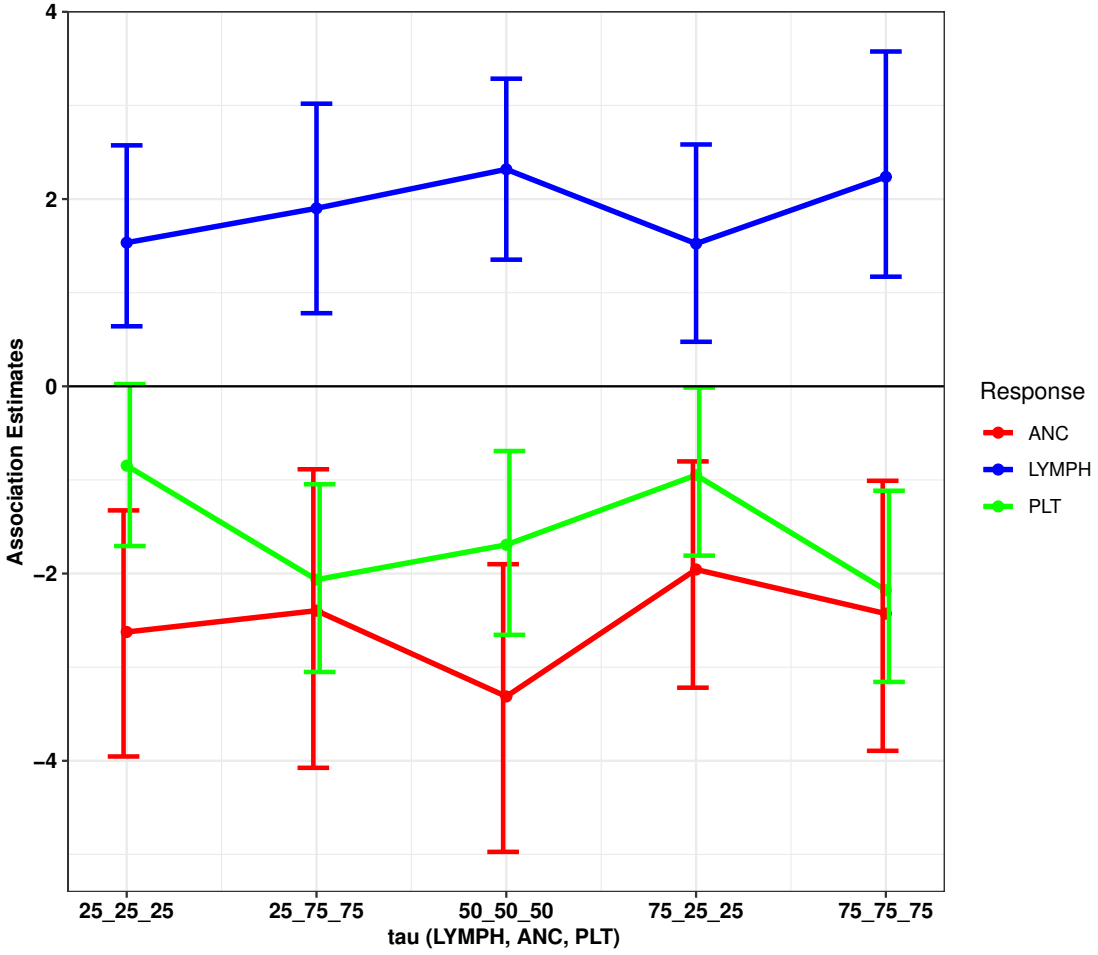


FIGURE 3.8: Estimate and 95% credible interval for the quantile-specific association coefficients of the three biomarkers to the event-time.

do not contain zeros. This indicates that an increase in the lymphocyte count results in a higher risk of relapse, which is intuitive. We also note that the estimated effect is the largest (in its absolute value) at the median level $\tau =$ (50,50,50).

The estimated association parameters for ANC and platelet count are all negative and the corresponding credible intervals do not contain zeros. This indicates that higher values of ANC and platelet count reduce the risk of relapse. The highest (negative) effect of ANC is observed at the median level, and for the platelet count the highest (negative) effect is observed at levels $\tau =$ (25,75,75) and $\tau =$ (75,75,75).

The estimated median non-relapse probabilities for the five quantile levels are shown in Figure 3.9. In this plot, the covariates with fixed effects are set to the median values, and the random effects (modeled as multivariate Brownian motion) are averaged over the subjects. It is noted that the median non-relapse probabilities are almost uniformly higher for $\tau =$ (25,75,75), i.e. when lower lymphocyte and higher ANC and platelet count are observed. On the other hand, the median non-relapse probabilities are almost uniformly lower for $\tau =$ (75,25,25), i.e. when a higher lymphocyte count and a lower ANC and platelet count are observed. This reassures that a higher lymphocyte count increases the relapse probability, and a higher ANC and a higher platelet count reduce it.

We note that the curve for $\tau =$ (25,25,25) dominates the curve for $\tau =$ (75,25,25), suggesting that a lower lymphocyte count improves the non-relapse probability. The combination $\tau =$ (75,75,75) has comparable non-relapse probabilities to that of $\tau =$ (25,75,75), suggesting that even if the lymphocyte counts are on the higher side, non-relapse probabilities can still be increased by ensuring higher ANC and platelet through medication. This is to say that no single model gives the best fit to this data which again defends our choice of quantile-specific joint modeling.

It is also to be noted that the median non-relapse probability in the end of the follow-up period for all quantile levels is at least 0.75. A relapse of ALL is always alarming since most of the relapses result in death. The government might go for a better healthcare system which can increase the non-relapse probability to 0.90 or higher.

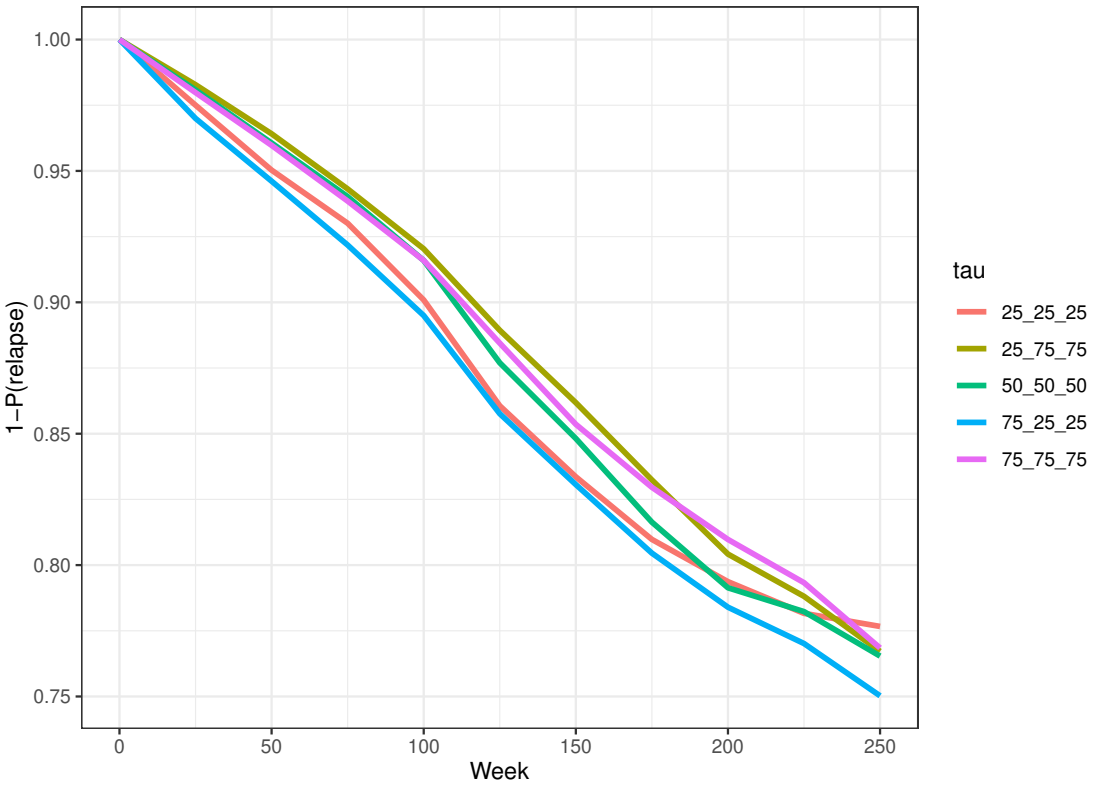


FIGURE 3.9: Median estimated non-relapse probability curves for different quantile combinations.

TABLE 3.14: Estimated correlation matrices (for three biomarkers) at different quantile levels. Lymp., ANC and Plt., respectively, denote lymphocyte count, neutrophil count and platelet count.

	τ=(25,25,25)			τ=(50,50,50)		
Outcome	Lymp.	ANC	Plt.	Lymp.	ANC	Plt.
Lymp.	1	0.682	0.588	1	0.590	0.432
ANC	-	1	0.830	-	1	0.580
Plt.	-	-	1	-	-	1
	τ=(25,75,75)			τ=(75,25,25)		
Lymp.	1	0.636	0.484	1	-0.029	0.045
ANC	-	1	0.562	-	1	0.808
Plt.	-	-	1	-	-	1
	τ=(75,75,75)					
Lymp.	1	0.664	0.504			
ANC	-	1	0.577			
Plt.	-	-	1			

3.4.2.4 Other findings

In Table 3.14, we report the estimated correlation matrices for three longitudinal biomarkers at different quantile levels. As expected the correlations vary from one quantile level to the other. At the lower and upper levels, i.e. at τ= (25,25,25) and at

τ= (75,75,75), and also at τ= (25,75,75) we observe moderate to high correlations among the outcomes. At the median levels, the correlations are moderate (between 0.4 and 0.6). We observe that correlation between ANC and platelet is greater than 0.8 when both the response are at the lower quantiles (i.e. for τ =(25,25,25) and τ= (75,25,25)). However, for $\tau = (75, 25, 25)$, the correlations between lymphocyte count and the other biomarkers are close to zero.

In Figure 3.10, we plot the estimated (marginal) quantiles for each biomarker separately at three different levels (i.e. 25, 50 and 75). We note that for all the three biomarkers the plots show non-decreasing trends indicating that we do not come across a quantile crossing issue in our analysis. However, this does not mean that our proposed model will never suffer from this issue. Biswas et al. (2020) [10] proposed a quantile smoothing method for handling the quantile crossing problem. The method can be used in our joint modeling as well.

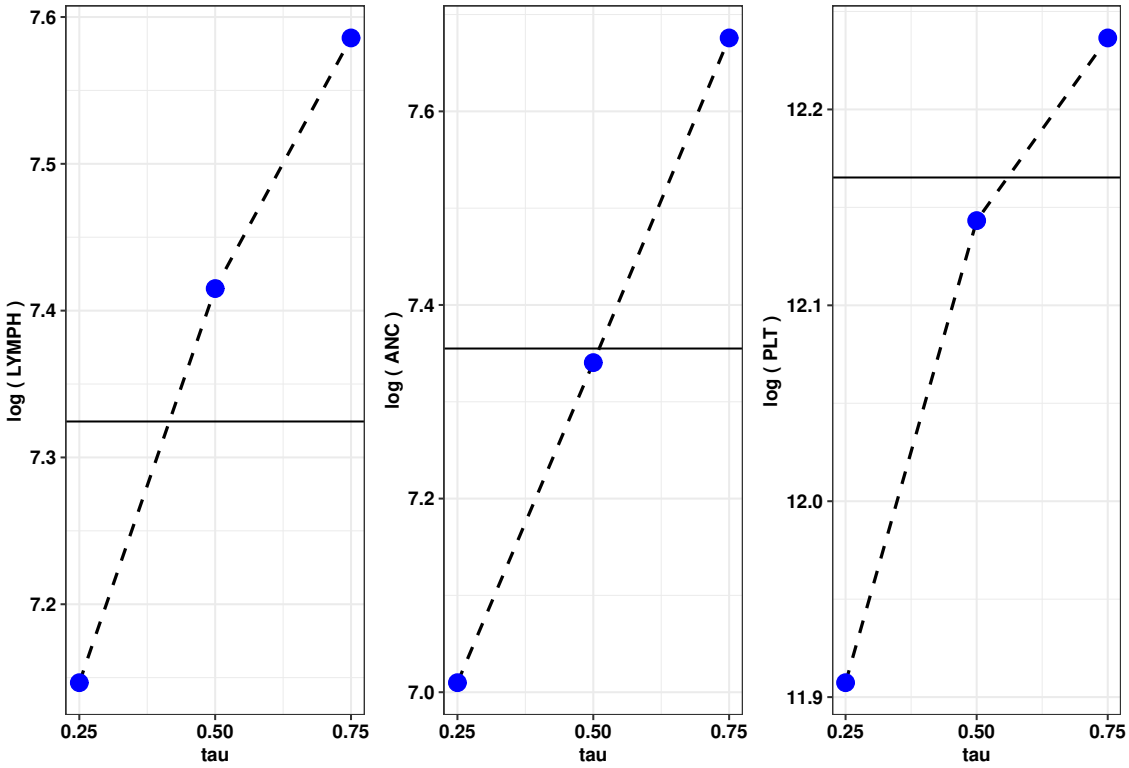


FIGURE 3.10: Estimated marginal quantiles for three different biomarkers for the ALL data analysis.

3.5 Simulation Study

We perform a simulation study for validating the proposed Bayesian quantile joint modeling. We simulate the longitudinal traits (Y_{ijk}) from the model given in equation (3.2), without the general effects of time. We consider two time-varying covariates,

i.e. $\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}]$, and three fixed covariates, i.e. $\mathbf{z}_i = [z_{i1}, z_{i2}, z_{i3}]$. All the covariates are generated from a standard normal distribution, and the subject-specific random effects are generated from a trivariate normal distribution with mean vector=0, and the covariance matrix Σ. The variance components in the matrix Σ are 2, 2.5,3; and the correlation between any two traits is fixed at 0.45. The random errors are generated either from a standard normal distribution or from ALD with different skewness parameters (τ_k).

For simulating the event-time, we use the model given in equation (3.3), with a constant baseline hazard. We consider ten longitudinal measurements for each subject, and then subjects are followed for the next fifteen time points. At T=25, subjects are censored. We consider the following three cases.

Case I: Random errors are generated from ALD with τ_k=0.25, $k = 1, 2, 3$ (right-skewed case).
Case II: Random errors are generated from ALD with τ_k=0.50, $k = 1, 2, 3$ (symmetric with heavy tails).
Case III: Random errors are generated from a standard normal distribution (symmetric).

For each of the three cases we simulate 100 datasets, and for each dataset we take 200 subjects. We fit the proposed quantile regression joint model (QRJM) and the traditional linear mixed joint model (LMJM) where the longitudinal process is modeled with linear mixed models and the event-time is modeled by a dynamic Cox PH model. We use MCMC for the parameter estimation.

For the joint modeling of longitudinal traits and event-time it is of interest to evaluate the discriminative capability of a model. We compute the area under the receiver operating characteristic curve (AUC) for different models. The AUC measures how efficiently a joint model discriminates the subjects for which a relapse occurred from the subjects with no relapse (Rizopoulos, 2016 [78]). Let $\pi_i(t + \Delta t|t)$ be the probability that for the i-th subject there is no relapse upto time $t + \Delta t$ given that it is event-free (no relapse) until time t. For any pair of subjects $[i, j]$ who are event-free until time t, the discriminative power of a model is assessed by computing AUC as below:

$AUC = P\left[\pi_i(t + \Delta t|t) < \pi_j(t + \Delta t|t)|(T_i \in (t, t + \Delta t]) \cap (T_j > t + \Delta t)\right]$, where T_i and T_j, respectively, denote the actual event-time for the i-th and the j-th subject. This means that for a fixed time-interval $(t, t + \Delta t]$ if a relapse occurs for the i-th subject but the j-th subject is event-free upto time $t + \Delta t$, then the model must assign a higher non-relapse probability to the j-th subject. We use this criterion to compare different models.

We consider t=10, and for three different values of Δt (i.e. Δt=5,8,12) we compute the true AUC (based on the simulated dataset) and the predicted AUC for three different models. Results (average AUC values from 100 datasets) are summarized

TABLE 3.15: AUC values for different models under different settings in the Simulation Study. Values are rounded upto two decimal places.

				Predicted AUC$(t, \Delta t)$		
Data Distribution	t	Δt	True AUC$(t, \Delta t)$	QRJM(τ=(25,25,25)	QRJM(τ=(50,50,50)	LMJM
ALD(τ_k=0.25)	10	5	0.83	0.82	0.81	0.73
		8	0.87	0.86	0.82	0.77
		12	0.91	0.90	0.87	0.82
ALD(τ_k=0.50)	10	5	0.86	0.83	0.86	0.82
		8	0.90	0.87	0.89	0.86
		12	0.92	0.89	0.91	0.88
$N(0,1)$	10	5	0.84	0.81	0.83	0.84
		8	0.88	0.85	0.86	0.87
		12	0.91	0.88	0.90	0.90

in Table 3.15. We note that when data are simulated from ALD with a specific skewness parameter, then QRJM (for that specific quantile level) gives the best prediction (highest AUC value). When we simulate data from a standard normal distribution, then the AUC values for QRJM with τ_k=0.5 (for $k = 1, 2, 3$) are quite comparable to those for LMJM. For all the other situations QRJM provides better prediction than LMJM. This study illustrates the usefulness of the proposed QRJM for a non-Gaussian setting.

3.6 Summary

In this chapter, we develop a Bayesian quantile-based joint model for multivariate longitudinal outcomes and event-time data. When the joint distribution of the biomarkers deviate from a multivariate normal distribution then a quantile-based regression model is used due to its robustness. Such models can also assess the evolution of different quantiles of the longitudinal outcomes, and their effects on the event-time. By exploiting a mixture representation of ALD (following Kozumi and Kobayashi, 2011 [50]) we develop computationally efficient Gibbs sampler algorithm for the proposed quantile-based joint model. Our analysis provides a complete picture on the covariate effects, and the complex association among the biomarkers, event-time and covariates. Our simulation studies also illustrate the effectiveness of the proposed model for a powerful Statistical inference.

There are, however, some limitations of our current work. Note that the effects of the drugs might differ from one age-group to the another, and therefore, an "age-group based" analysis might reveal some insights in the functioning of the drugs. Finally, there might be some latent classes in the dataset due to the difference in the evolution of the biomarkers, and the covariate effects might differ from one class level to the other. A latent-class joint model might be helpful for addressing this issue. In Chapter 4, we develop a Bayesian latent class model for jointly analysing multivariate longitudinal outcomes and event-time data.

Chapter 4

A latent class Bayesian joint model for multivariate longitudinal and event-time data

4.1 Preamble

Joint analysis of univariate or multivariate longitudinal outcomes and the time to the occurrence of one or more events of interest is an active research area over the last two decades. Joint models are quite popular in biomedical studies where a group of subjects are followed for a certain period of time, and the variables of interest are measured longitudinally, in addition to the event-time (if any). In a joint analysis, one can model the evolution of the longitudinal outcomes and the effects of those outcomes on the event-time. Additionally, the effects of the covariates on the evolution of the longitudinal process and the event-time can be assessed effectively in a joint analysis (Wang and Taylor, 2001 [96]; Fieuws and Verbeke, 2004 [30]; Rizopoulos and Ghosh, 2011 [79]; Das, 2016 [19]; Kundu et al., 2023 [52]).

Joint models are typically based on shared random effects or latent-class models. In the shared random effects models the longitudinal biomarkers are first modeled using (multivariate) linear mixed models, and then the event-times are modeled by the semi-parametric Cox proportional hazards (PH) model where the same random effects (used in modeling the longitudinal outcomes) are used as covariates in the PH model. Shared random effects capture the dependence between the evolution of the longitudinal outcomes and the event-time. Such models assume that the effects of the covariates are same across all the subjects conditional on the random effects (Henderson et al., 2000 [40]; Liu et al., 2008 [56]; Xu and Zeger, 2001 [100]; Zeng and Cai, 2005 [105]; and the references therein).

Joint latent-class models, on the other hand, assume that there are several subgroups in the dataset under consideration, and the evolution of the longitudinal outcomes and their effects on the event-time differ across different subgroups. Lin et

al. (2002) [55] developed a latent-class joint model with an application to longitudinal prostate-specific antigen readings and prostate cancer. Liu et al. (2015) [57] used similar latent-class model with an application to CPCRA study. More recently, Wong et al. (2022) [98] proposed a semi-parametric latent-class model for the joint analysis of multivariate longitudinal and event-time data. Proust-Lima et al. (2014) [69] gives a nice review on the recent developments on such latent-class joint modeling. Such models are based on finite mixture models for the longitudinal outcomes, and assess the group-specific effects of the outcomes on the event-time. Since the group information at the subject level is missing, expectation-maximization (EM) type algorithms (or the similar models in a Bayesian framework) are typically used for estimating the model parameters.

Our work in this chapter is motivated by the ALL dataset discussed in Section 1.4.1. However, similar to Chapter 3, we consider the lymphocyte count, the neutrophil count and the platelet count as the longitudinal biomarkers, and model them jointly along with the time-to-relapse (or the censoring time).

In latent-class joint modeling, typically the subjects are clustered with respect to their initial covariate values, and a multinomial logit model is used for such clustering (Wong et al., 2022 [98], and the references therein). Our approach in this work is more in the line of Putter et al. (2008) [73] where the goal is to distinguish the patients with distinct patterns in the longitudinal biomarkers. However, the results of the clustering based on evolution of multiple longitudinal biomarker are difficult to interpret, and computationally challenging as well. For our application, since the lymphocyte count plays the major role in the development and evolution of acute lymphocytic leukemia (ALL), we consider latent classes with different evolution patterns for the lymphocyte count only. A finite mixture linear mixed model is used for the lymphocyte count, whereas the neutrophil count and the platelet count are modeled by the traditional multivariate linear mixed models. For modeling the event-time using the Cox PH model we use the class-specific model, and the association among the three biomarkers are effectively considered in the event-time submodels.

In a Bayesian framework we estimate the model parameters using Markov Chain Monte Carlo (MCMC) algorithm, and compute the posterior probability for each individual to be assigned to each latent class. Our analysis finds two latent classes with distinct patterns in the evolution of lymphocyte, and we also compute the average non-relapse probability for each latent class. The trajectories for the neutrophil and platelet count for the latent classes are also shown for understanding the association among the three biomarkers.

The rest of the chapter is organized as follows. In Section 4.2 we provide a description of the dataset and specify the research goals. We also discuss the motivation for a joint latent-class modeling for this dataset in this section. In Section 4.3, the proposed models and the parameter estimation methods are discussed. The findings from the data analysis are summarized and discussed in Section 4.4. In Section 4.5,

we show the results from simulation studies which illustrates the effectiveness of the proposed modeling. Finally, some concluding remarks are given in Section 4.6.

4.2 Dataset and Motivation

The dataset for this analysis is the same as the one in Chapter 3. In this analysis, we consider the three longitudinal biomarkers, namely, lymphocyte count (LYM), neutrophil count (ANC) and platelet count (PLT) as response and the medicine dosage 6MP and MTx as time-varying covariates as before. In our dataset there were some missing observations in the biomarker values for some patients. The overall percentage of missingness for WBC, ANC and PLT counts were 32.04%, 4.83% and 4.80% respectively. In our analysis we deleted all the time-points with missing biomarker values and considered only those patients who had at least five non-missing biomarker values. This resulted in the total number of patients as 184 in the dataset. In addition to this all the time-points for patients in the treatment phase and follow-up phase were shifted to the same starting time for ease of comparison. The number of observations in the treatment phase varied from 5 to 60, with median number of visits being 28.

In the resulting data, 36% were females and the rest were males; the age range of the children varied in the interval [1, 17.5] with the median age being 4.45 years at presentation. About 28% of them have bulky disease (i.e., they have a cancerous mass with 10 cm or larger diameter) and 36% were affected by disease related to central nervous system (CNS). By the end of the study about 32% of the patients experienced a relapse. In our analysis we had to exclude some of the covariates, such as, Day 8 risk, Day 35 risk and MRD status from the list of fixed covariates as mentioned in Table 1.3, since these were measured at some fixed time point either during the treatment or on completion of the treatment. Since the goal of our analysis is to identify clusters based on the lymphocyte count in the presence of other responses, only the covariates that were measured at the beginning are considered in this analysis. Even though morphological remission for the subjects are measured at the end of their treatment phase, it is based on just the neutrophil and platelet counts, which we model for their respective mean trends later in our analysis. In addition to this, we also include another fixed covariate “Risk” which is a stratification made by doctors based on covariates observed before the treatment phase. All these time-invariant covariates for the 184 subjects are presented in Table 4.1.

TABLE 4.1: Summary statistics for the time-invariant covariates in the ALL dataset for 184 subjects.

Variable	Summary
Age at diagnosis	Min= 1, Q1=3, Median=4.45, Q3=7.808, Max= 17.5
WBC count at presentation	Min=100, Q1=5975, Median=17544, Q3= 46248, Max= 983500
Gender	Female: 36%, Male: 64%
Lineage	B cell: 85%, T cell: 15%
NCI risk group	High Risk: 36%, Standard Risk: 64%
Bulky disease	Yes: 28%, No: 69%, Unknown: 3%
CNS disease	Yes: 36%, No: 60%, Unknown: 4%
Risk	Good: 52%, Poor: 5%, T-cell: 2%, Other: 31%, Unknown: 10%
Risk at presentation	High Risk: 21%, Standard Risk: 45%, Intermediate Risk: 34%
Morphological Remission	Yes: 93%, No: 3%, Unknown: 4%

Figure 3.2 of Chapter 3 gives an idea that the relapse times might be affected by the biomarkers. We want to analyse the data (i) to detect the predictors which simultaneously influence the biomarkers and the relapse times, (ii) to detect the underlying subgroups exhibiting different patterns of the lymphocyte evolution over time, and (iii) to find how the association of the biomarkers and effects of the medicines vary across the latent subgroups. To answer these questions we implement a latent class Bayesian model for jointly analysing the biomarkers and the event-time.

4.3 Model and Methods

In our dataset, we have three biomarkers, namely, the (i) lymphocyte count (LYM) (ii) Neutrophil count (ANC) and (iii) Platelet count (PLT), which are longitudinal in nature. For stabilizing the variances of the response biomarkers we consider the log transformation (Kundu et al., 2023 [52]) and centered them at their respective median values. Let Y_{ijk} denote the k-th biomarker measured from the i-th patient at time t_{ij}, for $j = 1, 2, \ldots, \tau_i$, and $k = 1, 2, 3$. For the i-th patient we either observe the actual relapse-time T_i, or the censoring time C_i. Define s_i=$T_i \wedge C_i$, and $\delta_i = 1$, if $T_i < C_i$; (and 0, otherwise) and we consider (s_i, δ_i) as our event-time data.

4.3.1 Longitudinal Sub-models

Since ALL is caused mainly due to an uncontrolled growth of the lymphocyte count, it is of interest to distinguish the patients with significantly different patterns in the evolution of the lymphocyte count. Therefore, we consider a Bayesian latent-class model for the lymphocyte count, whereas for the neutrophil and the platelet count we use the traditional multivariate linear mixed models (Kundu et al., 2023 [52]) used for modeling the longitudinal outcomes.

4.3.1.1 Latent Class Model

Let Y_{ij1} denote the lymphocyte count for the i-th patient observed at time t_{ij}, and let G_i denote the latent subgroup the i-th patient belongs to. This is modeled as follows:

$$\begin{aligned} Y_{ij1}|(G_i = g) &= \mu_{ij1}^{(g)} + \epsilon_{ij1}^{(g)}; \\ \mu_{ij1}^{(g)} &= f_1^{(g)}(t_{ij}) + \boldsymbol{\beta}_1^{(g)\,T}\mathbf{x}_{ij} + \boldsymbol{\beta}_2^{(g)\,T}\mathbf{z}_i + a_i^{(g)} + b_i^{(g)}t_{ij}. \end{aligned} \tag{4.1}$$

In equation (4.1), the expected trajectory of lymphocyte for the g-th cluster at time t_{ij} is given by $\mu_{ij1}^{(g)}$, and $f_1^{(g)}(t) = \sum_{u=0}^{r} \eta_{1u}^{(g)} t^u$ represents the general effect of time as r-th degree polynomial. The effect of medicines, $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2})^T$ is given by $\boldsymbol{\beta}_1^{(g)}$, whereas $\boldsymbol{\beta}_2^{(g)}$ denotes the effect of fixed covariates $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^T$. Random effects $\boldsymbol{\alpha}_i^{(g)} = (a_i^{(g)}, b_i^{(g)})^T$, which contain the random intercepts $a_i^{(g)}$ and the random slopes $b_i^{(g)}$ are jointly normally distributed, i.e. $\boldsymbol{\alpha}_i^{(g)} \sim N_2(\mathbf{0}, \Sigma^{(g)})$. These effects capture the longitudinal dependence among the measurements for the same biomarker at different time points as well as the dependence among the biomarkers. Lastly, the measurement error for the g-th class, denoted by $\epsilon_{ij1}^{(g)}$, are assumed to be independently distributed as $N(0, \sigma^2_{1(g)})$.

4.3.1.2 Linear Mixed model for ANC and Platelet Counts

Let Y_{ij2}, Y_{ij3} denote the neutrophil and platelet count for the i-th patient observed at time t_{ij} ; $j = 1, \ldots, \tau_i$. We model these two outcomes as follows:

$$\begin{aligned} Y_{ijk} &= \mu_{ijk} + \epsilon_{ijk};\ k = 2, 3; \\ \mu_{ijk} &= f_k(t_{ij}) + \boldsymbol{\beta}_{1k}^T\mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^T\mathbf{z}_i + c_{ik} + d_{ik}t_{ij}. \end{aligned} \tag{4.2}$$

Similar to equation (4.1), the expected trajectory of the k-th response at time t_{ij} is given by μ_{ijk} in equation (4.2). Here, $f_k(t) = \sum_{u=0}^{r} \eta_{1uk} t^u$ serves similar purpose as $f_1^{(g)}(t)$ and the value of r is selected based on some model selection criteria. The effects of medicines and the fixed covariates on the k-th response are given by $\boldsymbol{\beta}_{1k}$ and $\boldsymbol{\beta}_{2k}$, respectively. Random effects $\boldsymbol{\gamma}_i = (c_{i2}, c_{i3}, d_{i2}, d_{i3})^T$ capture dependence between longitudinal measurements within and across the responses by assuming that $\boldsymbol{\gamma}_i \sim N_4(\mathbf{0}, \Sigma_{23})$. The random errors for the k-th response are given by $\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_k^2)$.

4.3.2 Event-time Sub-model

The relapse-time of a patient is possibly linked to the longitudinal outcomes, and the fixed predictors. We use a Cox-PH model to study the association of the longitudinal biomarkers to that of the observed relapse-times, s_i. It is intuitive that the hazard rate of the subjects will be cluster-specific where the within cluster hazard model

resemble largely as the hazard model in Rizopoulos (2016) [78].
The expected longitudinal trajectories of lymphocyte, neutrophil and platelet counts at time t are given by $\mu_{i1}^{(g)}(t), \mu_{i2}(t)$ and $\mu_{i3}(t)$, respectively, where $\mu_{i1}^{(g)}(t) = f_1^{(g)}(t) + \boldsymbol{\beta}_1^{(g)\,T}\mathbf{x}_i(t) + \boldsymbol{\beta}_2^{(g)\,T}\mathbf{z}_i + a_i^{(g)} + b_i^{(g)}t$, and $\mu_{ik}(t) = f_k(t) + \boldsymbol{\beta}_{1k}^T\mathbf{x}_i(t) + \boldsymbol{\beta}_{2k}^T\mathbf{z}_i + c_{ik} + d_{ik}t;\ k = 2, 3$. The symbols hold similar meaning as in equations (4.1) and (4.2), and $\mathbf{x}_i(t)$ is the quantity of 6MP and MTx administered at time t for the i-th patient. To serve our inference goals we take the association of all 3 responses to be cluster-specific. Note that ANC and platelet counts are not modeled at the (latent) cluster level, but might have varied effects on the hazard rates across the clusters. Therefore, we consider the cluster-specific PH model as follows:

$$\lambda_i(t)|(G_i = g) = \lambda_0^{(g)}(t)\exp\left(\psi_1^{(g)}\mu_{i1}^{(g)}(t) + \sum_{k=2}^{3}\psi_k^{(g)}\mu_{ik}(t) + \boldsymbol{\theta}^{(g)T}\mathbf{z}_i\right). \tag{4.3}$$

In equation (4.3), $\lambda_0^{(g)}(t)$ denotes the baseline hazard for the g-th class, which is modeled by $log(\lambda_0^{(g)}(t)) = \nu_0^{(g)} + \sum_{l=1}^{Q}\nu_l^{(g)}B_q(t, \boldsymbol{\zeta})$, where, $B_q(t, \boldsymbol{\zeta})$ is q-th basis function of B-splines with knots $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_Q)^T$. The knots are obtained by considering a large number of evenly spaced quantiles of the observed relapse-times (as used in JMBayes package) and then penalize the B-spline coefficients by considering suitable prior distributions (e.g., Laplace prior, Horseshoe prior etc.). The class-specific associations of mean lymphocyte, ANC and platelet counts are given by $\psi_1^{(g)}, \psi_2^{(g)}$ and $\psi_3^{(g)}$ respectively. Lastly, the class-specific baseline coefficients for the baseline covariates $\mathbf{z}_i$ is given by $\boldsymbol{\theta}^{(g)}$.

4.3.3 Joint Likelihood and Bayesian Estimation

Let $\boldsymbol{\Theta}^{(g)}$ be the set of model parameters specific to cluster g; $g = 1, \ldots, G$, and $\boldsymbol{\Theta}_{23}$ be the set of model parameters from equation (4.2). Then the joint likelihood for the set of parameters $\boldsymbol{\Theta} = (\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(G)}, \boldsymbol{\Theta}_{23})$ is given as follows:

$$L(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{s}, \boldsymbol{\delta}) = \prod_{i=1}^{n}\sum_{g=1}^{G}\left[\left(\prod_{j=1}^{\tau_i} f(Y_{ij1}|\boldsymbol{\alpha}_i^{(g)}, \boldsymbol{\Theta}^{(g)})\right) \times \left(\lambda_i(s_i|G_i = g, \boldsymbol{\alpha}_i^{(g)})\right)^{\delta_i} \times \exp\left(-\int_0^{s_i}\lambda_i(t|G_i = g, \boldsymbol{\alpha}_i^{(g)})dt\right) f(\boldsymbol{\alpha}_i^{(g)}|\Sigma^{(g)})P(G_i = g)\right] \times L_i(\boldsymbol{\Theta}_{23}|\boldsymbol{Y}_{i2}, \boldsymbol{Y}_{i3}). \tag{4.4}$$

In equation (4.4), $L_i(\boldsymbol{\Theta}_{23}|\boldsymbol{Y}_{i2}, \boldsymbol{Y}_{i3})$ gives likelihood for equation (4.2), and

$$L_i(\boldsymbol{\Theta}_{23}|\boldsymbol{Y}_{i2}, \boldsymbol{Y}_{i3}) = \left[\prod_{k=2}^{3}\prod_{j=1}^{\tau_i} f(Y_{ijk}|\boldsymbol{\gamma}_i)\right] \times f(\boldsymbol{\gamma}_i|\Sigma_{23}).$$

The $f(.)$ in equation (4.4) are the normal densities as in equations (4.1) and (4.2), respectively. The g-th class specific hazard rate $\lambda_i(.)$ in equation (4.4) consists of the both class-specific parameters $\boldsymbol{\Theta}^{(g)}$, and global parameters $\boldsymbol{\Theta}_{23}$.

We use the likelihood given by equation (4.4), and by considering appropriate prior distributions on such parameters we draw samples from the joint posterior distribution. All our inferences are based on the joint posterior distribution. We implement MCMC algorithm for estimating the model parameters, all our computations are performed using *JAGS* in *R*.

4.3.3.1 Prior and Joint Posterior Distribution

We mostly use the diffuse priors for the model parameters, similar to the priors used in Kundu et al. (2023) [52], Rizopoulos (2016) [78]. The components of η and β in equations (4.1) and (4.2) follow $N(0, 1000)$. For the g-th class covariance matrix for random effects of lymphocyte ($\Sigma^{(g)}$) we consider an Inverse Wishart ($\boldsymbol{I}_2, 3$) prior distribution, and for the covariance matrices corresponding to the random effects of neutrophil and platelet counts we also consider Inverse Wishart priors, i.e. $\Sigma_{23} \sim IW(\boldsymbol{I}_4, 5)$. For the g-th class association coefficients $\psi_1^{(g)}, \psi_2^{(g)}, \psi_3^{(g)}$ and for elements of $\boldsymbol{\theta}^{(g)}$ we consider N(0,1000) prior distribution. For our analysis, a penalized prior is applied on 15-knot cubic B-spline coefficients $\boldsymbol{\nu}^{(g)} = (\nu_0^{(g)}, \ldots, \nu_Q^{(g)})^T$, similar to Rizopoulos (2016) [78]. For the g-th class error precision of lymphocyte count and for the error precision of neutrophil and platelet count we consider Gamma (0.001,0.001) prior distribution. Finally, for the class probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^T$ we consider a flat prior, that is $\boldsymbol{\pi} \sim$ Dirichlet$(1, 1, \ldots, 1)$. Let $\pi(\boldsymbol{\Theta})$ be the joint prior distribution for the complete set of parameters $\boldsymbol{\Theta}$. Then the joint posterior distribution is given as follows:

$$\pi(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{s}, \boldsymbol{\delta}) \propto L(\boldsymbol{\Theta}|\boldsymbol{Y}, \boldsymbol{s}, \boldsymbol{\delta}) \times \pi(\boldsymbol{\Theta}).$$

The posterior probability (p_{ig}) that the i-th patient will belong to class g, is given as follows:

$$p_{ig} = \frac{L_{i,g} \times \pi_g}{\left[\sum_{c=1}^{G} L_{i,c} \times \pi_c\right]}, \tag{4.5}$$

with

$$L_{i,g} = \left[\left(\prod_{j=1}^{\tau_i} f(Y_{ij1}|\boldsymbol{\alpha}_i^{(g)}, \boldsymbol{\Theta}^{(g)})\right) \times \left(\lambda_i(s_i|G_i = g, \boldsymbol{\alpha}_i^{(g)})\right)^{\delta_i} \times \exp\left(-\int_0^{s_i} \lambda_i(t|G_i = g, \boldsymbol{\alpha}_i^{(g)})dt\right) \times f(\boldsymbol{\alpha}_i^{(g)}|\Sigma^{(g)})\right].$$

It is interesting to note that $\boldsymbol{\Theta}_{23}$ being the set of parameters that contributes to the mean longitudinal trajectory of neutrophil and platelet count due to equation (4.2), also influences the value of p_{ig} through their impact on the hazard rate and hence the class specific non-relapse probabilities in conformity with equation (4.3). We can use p_{ig} to identify the class in which the i-th patient is most likely to belong, i.e., the posterior estimate of the class indicator is given by, $\hat{G}_i = \underset{g}{\operatorname{argmax}}\, \hat{p}_{ig}$, where $\hat{p}_{ig}$ is obtained by plugging in the posterior estimates of $\boldsymbol{\Theta}^{(g)}$, $\boldsymbol{\Theta}_{23}$ and $\boldsymbol{\pi}$ in equation (4.5).

4.4 Data Analysis

4.4.1 Computational Details

We use MCMC iterations (based on Gibbs sampler and Metropolis-Hastings Algorithm) for estimating the model parameters. We run 20,000 iterations and discard the first 5,000 iterations as "burn-in", and thin the chains by saving every 10-th iteration, for each of the 3 independent chains. The entire computation is done in *JAGS 4.3.0*. The trace plots, cumulative mean plots and density plots for the model parameters indicate the convergence of the chains. In addition, we compute scale reduction factor (Brooks and Gelman, 1998 [12]) for assessing convergence in chains and the computed scale reduction factors were all less then 1.2. Figure 4.1 - 4.3 shows the plots for convergence for some of the model parameters.

While performing the computation we come across a problem with the Inverse Wishart prior that was considered for the covariance matrices mentioned in equations (4.1) and (4.2). The error is "*Unable to find appropriate sampler*". This happens because the determinant of the covariance matrices mentioned above gets small enough and beyond the tolerance limit of JAGS which makes them not-invertible. To evade this problem we consider $\left(\Sigma^{(g)}\right)^{-1} = \sum_{u=1}^{3} \Delta_u^{(g)} \left(\Delta_u^{(g)}\right)^T + 0.001\boldsymbol{I}_2$, and $\left(\Sigma_{23}\right)^{-1} = \sum_{u=1}^{5} \chi_u \left(\chi_u\right)^T + 0.001\boldsymbol{I}_4$, where, $\Delta_u^{(g)} \overset{iid}{\sim} N_2(\mathbf{0}, \boldsymbol{I}_2)$ and $\chi_u \overset{iid}{\sim} N_4(\mathbf{0}, \boldsymbol{I}_4)$. Notice that 0.001 times the Identity matrices are there to ensure the positive definiteness of the precision matrices. All computations were done in *R*, in a Windows 10, i7 processor machine it takes nearly 36 hours for the complete analysis.

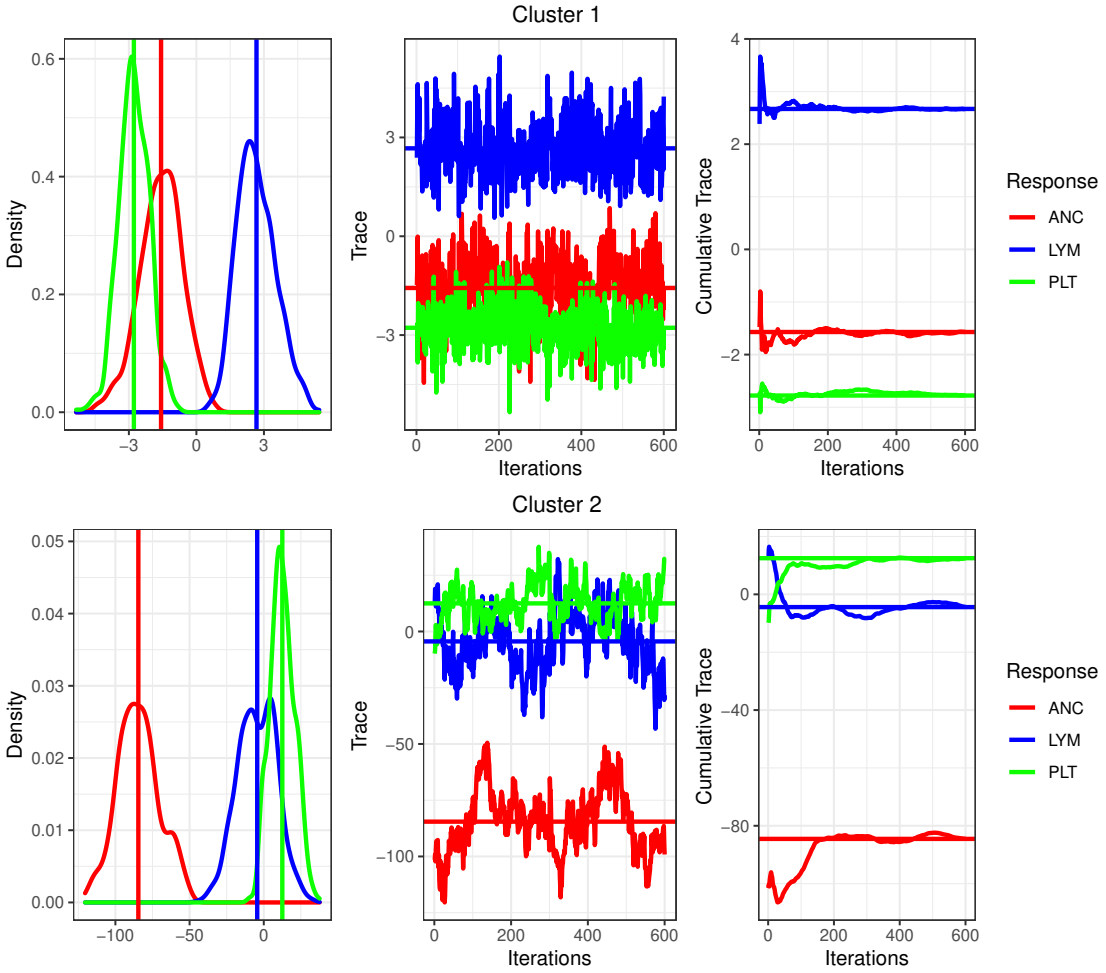


FIGURE 4.1: Density plot, trace plot and cumulative trace plot for the cluster-specific association parameters in the event-time sub model

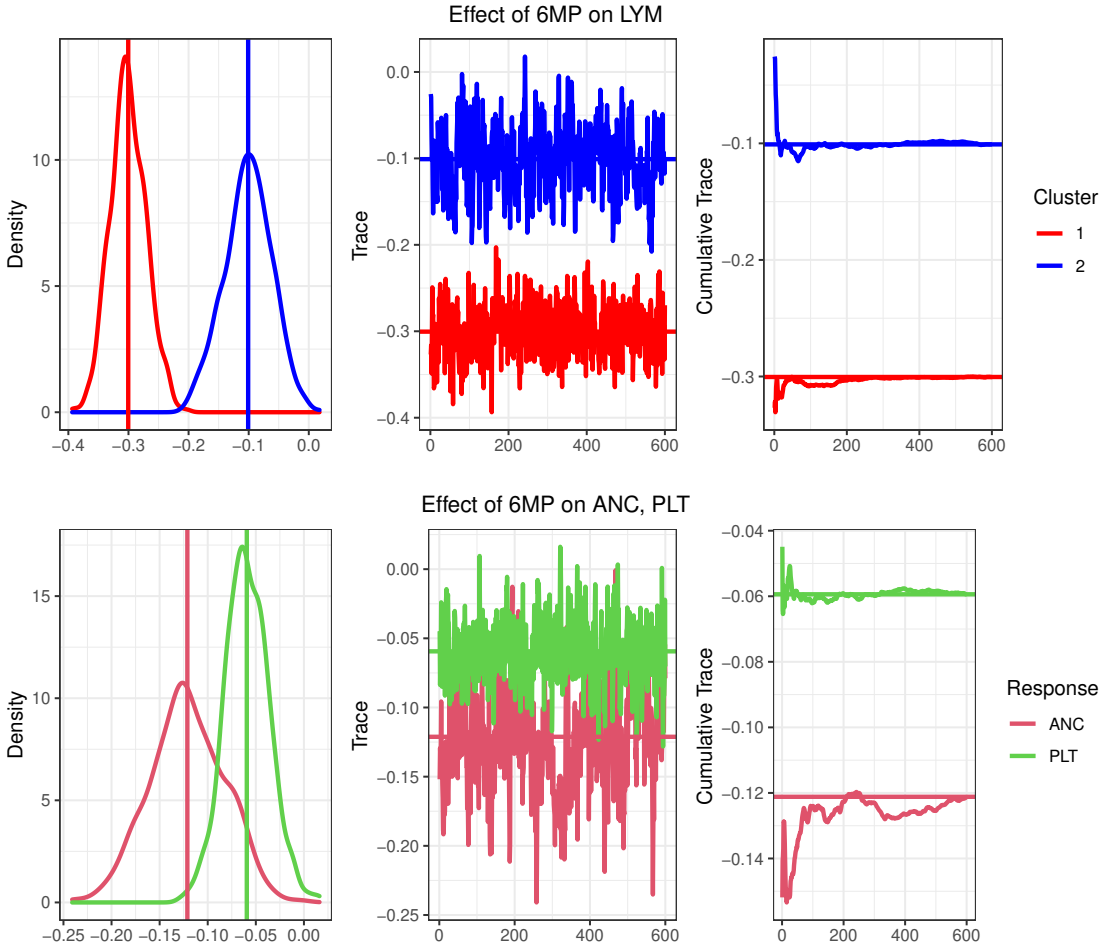


FIGURE 4.2: Density plot, trace plot and cumulative trace plot for the coefficients of medicine 6MP in equation (4.1) and (4.2)

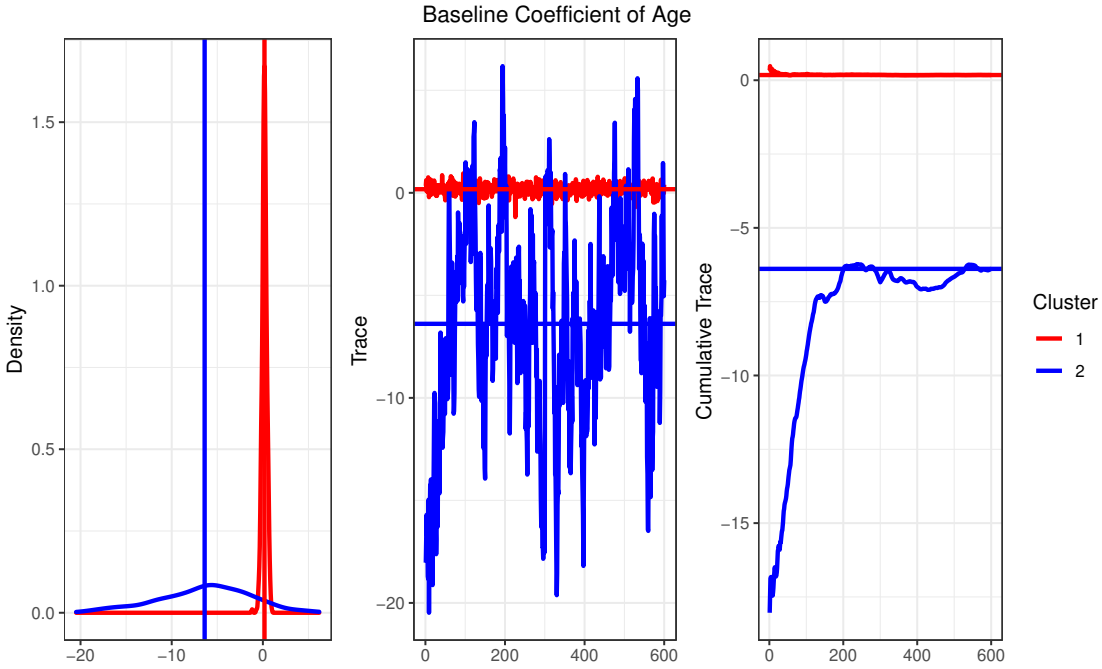


FIGURE 4.3: Density plot, trace plot and cumulative trace plot for the coefficients of fixed covariate Age in the event-time model

4.4.2 Optimal Number of Latent classes

For selecting the optimal number of latent classes (G) a series of models for different values G are to be tested. One can look at the size of the smallest cluster produced by a model and if the "best fit" model results in G classes, and the size of its smallest class is < 5, then the optimal number of classes is taken as $G - 1$.

We use the usual model selecting criteria such as AIC, BIC, DIC for selecting the optimal number of latent classes. The main problem with using DIC is that if the size of the smallest class is negligibly small (say < 5 subjects i.e., about 2.7%), then the value of DIC does not change much for model with total class $G - 1$ than the model with G classes. But the other criteria such as AIC and BIC, along with taking into account the likelihood also penalize the total number of estimated parameters. Since the total number of estimated parameters increase as we consider models with higher number of total classes, we rely on the AIC and BIC values. The BIC criteria was also used in Lin et al. (2002) [55], Liu et al. (2015) [57], Muthen and Shedden (1999) [61] for selecting optimal number of latent classes.

In Table 4.2 we observe that the minimum class size is lower than the limit 2.7% for $G = 3$, whereas, that for $G = 2$ is substantially bigger than the limit. Although the DIC is the lowest for model with $G = 3$, in terms of AIC and BIC the model with $G = 2$ gives lower values than the model with $G = 3$. Thus, we consider G=2 as the

optimal number of classes in our analysis. Note that DIC reported in this Chapter is not scaled by the number of subjects, n unlike in the Chapters 2 and 3.

TABLE 4.2: Model log-Likelihood, DIC, AIC, BIC and size of smallest class based on $\hat{G}_i$ values for selecting the optimal value of G in ALL data analysis.

G	log-Likelihood	DIC	AIC	BIC	minimum cluster
2	-8232.647	18224.036	19709.293	24923.919	27.174%
3	-8186.962	18185.209	20455.924	27017.608	1.630%

4.4.3 Findings

In Tables 4.3 and 4.4, we summarise the estimated coefficients and 95% Bayesian credible intervals (based on the MCMC iterations) for the longitudinal sub-model. We consider a covariate to be significant if their 95% CI does not contain zero (Das, 2016 [19]). For the optimal degree of polynomial f in equations (4.1) and (4.2) we compute the DIC values with r=1,2, and 3; and the smallest DIC value was obtained for r=2. Thus, we consider a quadratic function of time for the general effect f.

We notice that effect of medicine 6MP is significant and negative for both the classes (of lymphocyte), and also in overall trend for ANC and platelet count. On the other hand, MTx is not significant for class 1 (of lymphocyte) but is negatively significant for the class 2, and has significant positive effect for ANC whereas not insignificant for the platelet count. Except for the covariate 'Risk', no fixed covariate was significant in class 1 of lymphocyte, while in class 2 the covariates age, gender and bulky disease are significant with negative effects. For the ANC, the covariates lineage, bulky disease, risk at presentation and morphological remission are significant, but for platelet count only morphological remission turned out to be significant.

In Table 4.5 we summarise the degree of association of the mean longitudinal trends to the event-time, and class-wise effects of the time-invariant covariates on the event-time. From this table we notice that the association effect of lymphocyte count in class 1 is significant and positive, but in class 2 it is insignificant. In class 1 association of neutrophil count is insignificant, but for platelet count it is significant and negative. But association of nuetrophil count in class 2 is significant and negative with a much higher magnitude compared to the rest of association estimates, while the platelet count association effect is marginally insignificant. This means that the non-relapse probability of the subjects that fall in class 1 can be increased by lowering lymphocyte count and increasing platelet count, but for subjects in class 2 only neutrophil count is to be reduced to increase non-relapse probability. Even though by increasing the 6MP dosage for the subjects in class 1 the platelet count is reduced, but it will reduce the lymphocyte count at a much higher rate such that their combined effect will reduce the hazard, and will increase the non-relapse probability. On the other hand, change in the MTx dosage for subjects in class 1 does not alter the hazard

rate much. Thus, we conclude that if the subjects in class 1 were given a high dosage of 6MP, then they would have a higher chance of non-relapse. We further investigate that due to the highly negative effect of neutrophil count on hazard for the subjects in class 2, a lower dosage of 6MP and a high dosage of MTx will result in a higher neutrophil count, and that increases the non-relapse probability.

Covariates such as lineage and bulky disease have significant effect negative effects for both classes in the event-time model, where as WBC at presentation has significant but has opposite effects in the class. The effect of covariate 'Risk' is significant and positive for class 1, but it is insignificant for class 2. In class 2, morphological remission and risk at presentation are negatively and positively significant, but they are insignificant for class 1.

In general it is observed from Table 4.5 that 95% CI of the fixed covariates and association coefficients in class 2 is larger than their counterparts in class 1. This is probably because the size of class 2 is much smaller than that of class 1. We observed the size of class 2 to be 27.17%. In Figure 4.4 we plot $\hat{p}_{i1}$, the estimated value from equation (4.5) for $g = 1$ where the i-th subject is assigned to class 1 if $\hat{p}_{i1} > 0.5$. It is clear from this plot that the subjects are well separated between two classes. Only 4 subjects lie very close to the 0.5 line, and in fact they lie with in the boundary lines of 0.45 and 0.55 respectively.

It is clear from Figure 4.5 that the mean lymphocyte count in class 1 is higher and grows more rapidly than that in class 2. The difference between these two curves change much in the latter part of the treatment phase. Based on Figure 4.6, we observe that subjects in class 1 have a lower (average) rate of non-relapse than class 2. For a better understanding, one needs to look at Figure 4.7, which shows the class-specific mean longitudinal curves for the neutrophil and platelet count. Since we did not consider the class-specific trajectories for these two biomarkers we get the similar trends in the mean trajectories for the two latent classes. However, for the neutrophil counts the curve for class 1 is consistently above the curve for class 2; and for the platelet count we observe a reverse trend. Overall, a higher mean curve for the lymphocyte and a lower mean curve for the platelet count reduce the average non-relapse probability for this class.

We observed that the Kaplan-Meier (K-M) (Kaplan and Meier, 1958 [47]) non-relapse probability curves for the latent classes in Figure 4.8 are similar to the estimated ones in Figure 4.6. However, one cannot expect the estimated non-relapse probabilities to be just the smoothed out versions of the K-M plots, since the class-specific non-relapse probabilities are heavily influenced by the expected longitudinal responses through the hazard rates. In general, it easy to observe that two latent classes are significantly different from each other with respect to the effects of the medicines and other covariates, as well as the mean non-relapse probabilities. A simple joint model (without latent classes) would fail to provide the insights that the proposed latent-class model provides us.

TABLE 4.3: Estimated coefficients and 95% credible interval for the covariates in the latent classes of Lymphocyte in longitudinal sub-model.

Covariate	Lymphocyte Class 1		Lymphocyte Class 2	
	Estimate	95% CI	Estimate	95% CI
6MP dose	-0.300	(-0.354,-0.239)	-0.101	(-0.181,-0.03)
MTx dose	0.027	(-0.039,0.097)	-0.095	(-0.174,-0.019)
Age at diagnosis	0.037	(-0.031,0.105)	0.113	(0.021,0.2)
Gender	0.016	(-0.061,0.089)	0.132	(0.019,0.251)
Lineage	-0.006	(-0.125,0.122)	-0.087	(-0.243,0.093)
WBC at presentation	-0.005	(-0.038,0.025)	0.051	(-0.002,0.106)
NCI Risk group	-0.038	(-0.174,0.078)	0.096	(-0.071,0.237)
Bulky disease	-0.036	(-0.116,0.041)	0.155	(0.044,0.262)
CNS disease	0.059	(-0.011,0.126)	-0.041	(-0.139,0.076)
Risk	-0.029	(-0.058,-0.003)	0.027	(-0.016,0.077)
Risk at presentation	0.006	(-0.044,0.059)	-0.068	(-0.172,0.003)
Morphological remission	0.110	(-0.017,0.251)	0.046	(-0.222,0.274)

TABLE 4.4: Estimated coefficients and 95% credible interval for the covariates for responses Neutrophil and Platelet counts in longitudinal sub-model.

Covariate	Neutrophil count		Platelet count	
	Estimate	95% CI	Estimate	95% CI
6MP dose	-0.121	(-0.196,-0.05)	-0.059	(-0.104,-0.013)
MTx dose	0.184	(0.097,0.26)	0.039	(-0.003,0.086)
Age at diagnosis	-0.023	(-0.075,0.028)	-0.049	(-0.117,0.028)
Gender	-0.017	(-0.091,0.06)	-0.024	(-0.135,0.069)
Lineage	0.153	(0.037,0.286)	0.023	(-0.153,0.245)
WBC at presentation	-0.019	(-0.053,0.014)	-0.015	(-0.054,0.031)
NCI Risk group	0.005	(-0.073,0.099)	0.125	(-0.013,0.246)
Bulky disease	-0.101	(-0.173,-0.038)	-0.085	(-0.223,0.008)
CNS disease	-0.006	(-0.088,0.063)	-0.007	(-0.094,0.087)
Risk	-0.008	(-0.034,0.015)	-0.010	(-0.049,0.026)
Risk at presentation	0.068	(0.02,0.121)	-0.034	(-0.113,0.044)
Morphological remission	0.300	(0.199,0.46)	0.301	(0.111,0.489)

TABLE 4.5: Estimated coefficients and 95% credible interval for the association parameters and baseline covariates in the latent classes in event-time sub-model.

Covariate	Class 1		Class 2	
	Estimate	95% CI	Estimate	95% CI
Lymphocyte count	2.670	(1.232,4.542)	-4.411	(-29.583,20.664)
Neutrophil count	-1.571	(-3.664,0.168)	-84.529	(-113.07,-56.702)
Platelet count	-2.778	(-4.001,-1.424)	12.497	(-1.967,28.032)
Age at diagnosis	0.176	(-0.329,0.668)	-6.389	(-17.264,2.734)
Gender	0.424	(-0.41,1.213)	5.209	(-6.564,17.415)
Lineage	-1.952	(-3.274,-0.821)	-24.808	(-46.617,-9.036)
WBC at presentation	0.376	(0.005,0.759)	-8.891	(-15.123,-2.078)
NCI Risk group	0.364	(-0.585,1.303)	0.941	(-10.413,13.983)
Bulky disease	-0.882	(-1.522,-0.242)	-23.543	(-36.591,-11.137)
CNS disease	0.457	(-0.272,1.223)	6.415	(-4.43,17.282)
Risk	0.425	(0.178,0.672)	-0.855	(-6.235,4.866)
Risk at presentation	-0.078	(-0.542,0.386)	17.161	(9.009,26.087)
Morphological remission	-0.106	(-1.901,1.568)	-34.601	(-55.497,-11.94)

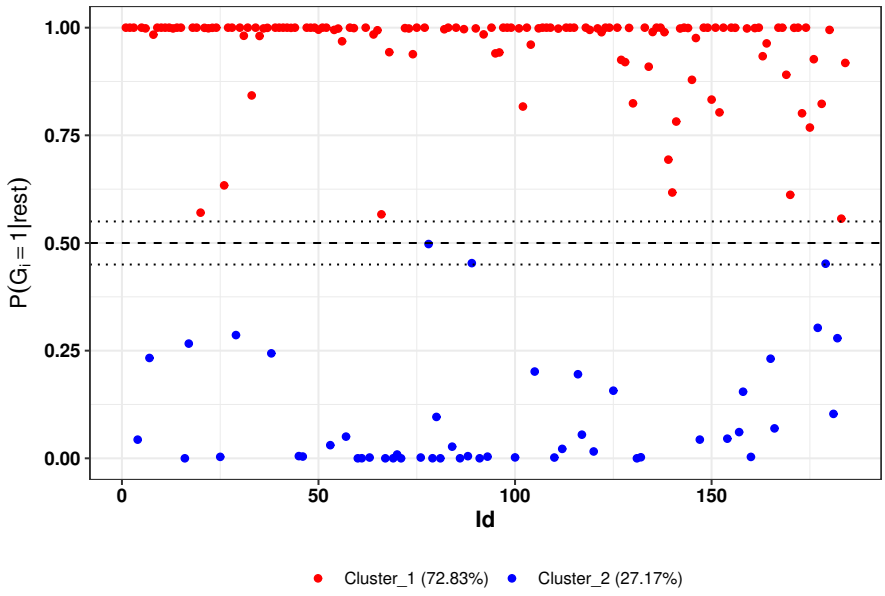


FIGURE 4.4: Posterior inclusion probabilities of class 1 (i.e., for all subjects in ALL study.

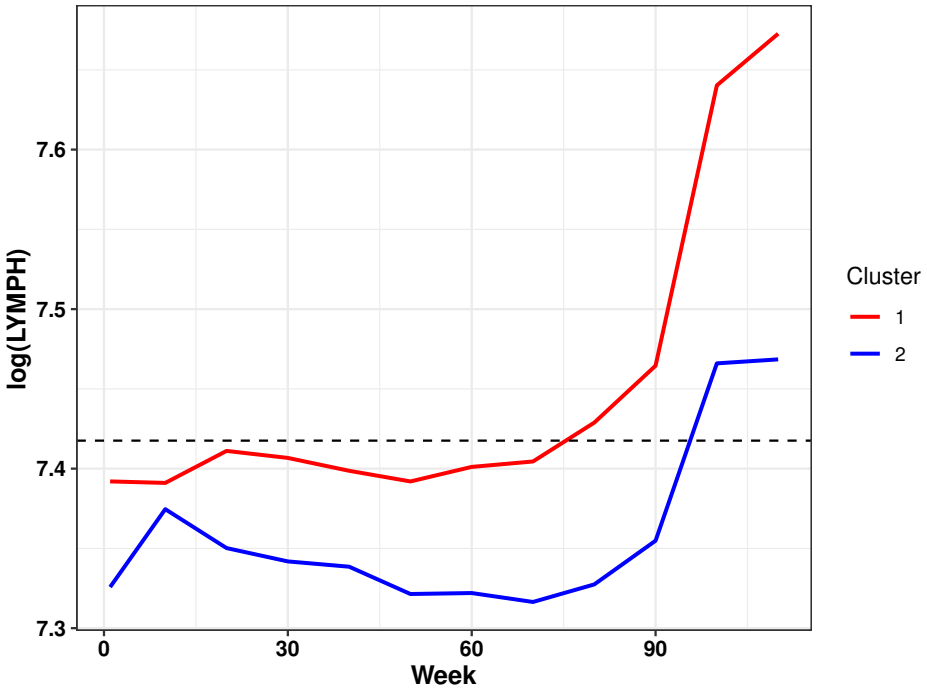


FIGURE 4.5: Longitudinal trajectories of latent clusters of Lymphocyte count

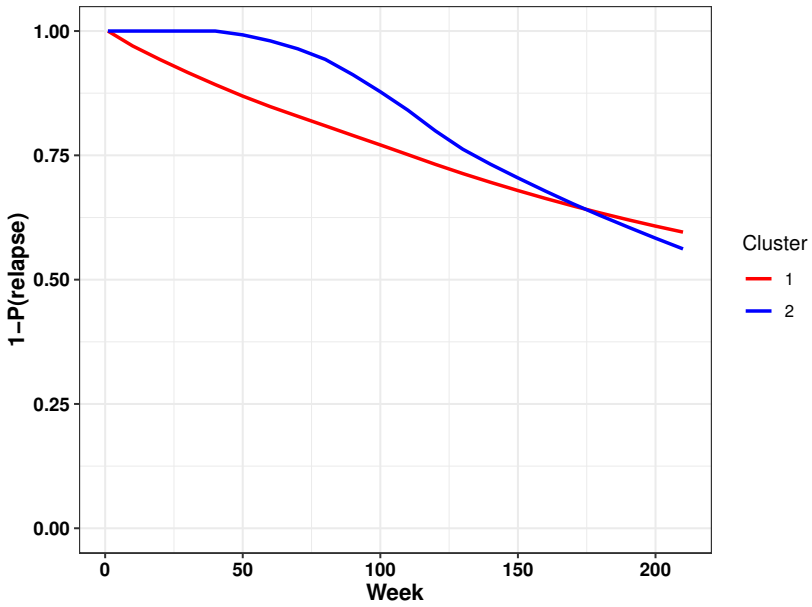


FIGURE 4.6: Non relapse probability curves for the latent clusters

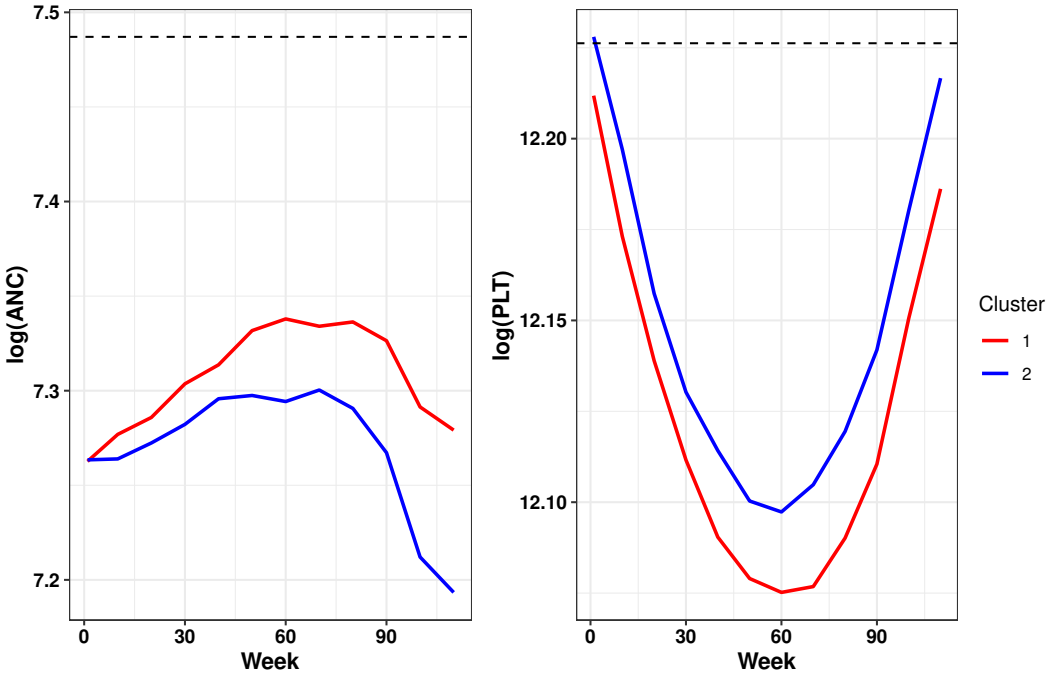


FIGURE 4.7: Longitudinal trajectories of Neutrophil and Platelet counts according to latent clusters. The dashed lines representing the mean responses in respective plots.

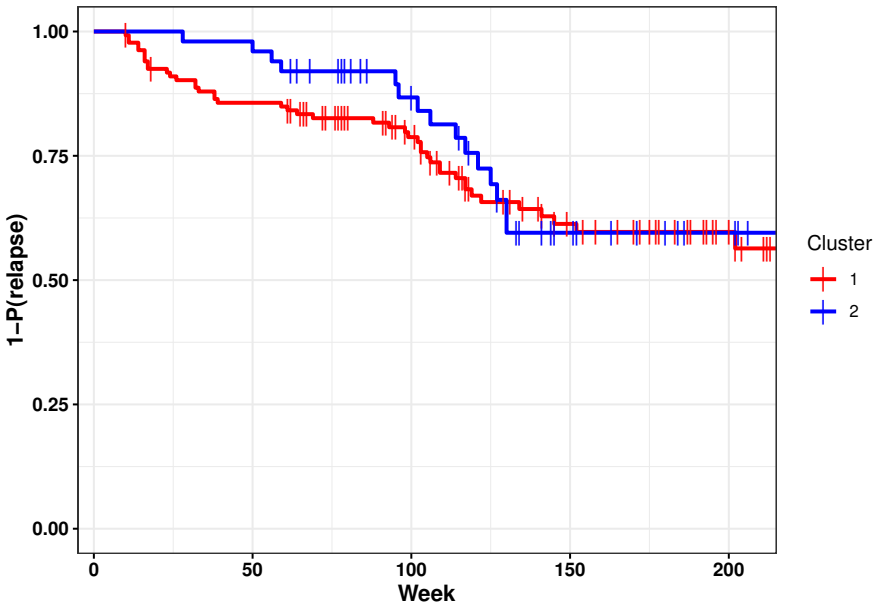


FIGURE 4.8: Kaplan Meier plots for the two latent groups, with the small vertical intercepts representing the censoring time

4.5 Simulation Studies

For assessing the practical usefulness and discriminative power of the proposed latent-class model we perform a simulation study. We simulate data for 200 subjects similar to ALL Chemotherapy dataset, for which we measure three longitudinal outcomes for the first 10 time points, and then they are followed for the next 15 time points. At time T=25, the subjects are censored. We measure two time-dependent covariates, and three time-invariant covariates from each subject.

We consider two latent-clusters with respect to the first longitudinal outcome, and generate data for that particular outcome using equation (4.1) with G=2. Then, we use the models in equation (4.2) for simulating the two other longitudinal biomarkers. Regression coefficients are chosen based on the estimated coefficients from the ALL data analysis. Finally, we use the class-based PH model (given in equation (4.3)) for simulating the event-time data. The regression coefficients for the PH model are chosen such that we observe the event-time for nearly 30% subjects. We simulate 100 replicates of the dataset (each containing 200 patients).

Once the datasets are simulated, we use three competing models for assessing their relative effectiveness. First, we consider a traditional joint model where we use linear mixed models for modeling each longitudinal outcomes, and then use a Cox PH model for the event-time with the mean longitudinal outcomes are taken as covariates and the random effects are shared. Specifically, the longitudinal outcomes are modeled as follows:

$$\begin{aligned} Y_{ijk} &= \mu_{ijk} + \epsilon_{ijk};\ k = 1, 2, 3; \\ \mu_{ijk} &= f_k(t_{ij}) + \boldsymbol{\beta}_{1k}^T \mathbf{x}_{ij} + \boldsymbol{\beta}_{2k}^T \mathbf{z}_i + c_{ik} + d_{ik} t_{ij}, \end{aligned} \tag{4.6}$$

where the symbols are similar to the model in equation (2). And then for modeling the event-time we use the following Cox PH model:

$$\lambda_i(t) = \lambda_0(t) \exp\left(\sum_{k=1}^{3} \psi_k \mu_{itk} + \boldsymbol{\theta}^T \mathbf{z}_i \right). \tag{4.7}$$

We refer to this model as Model 1. Similar to the joint likelihood in equation (4.4), we can write the joint likelihood for this model, and estimate the model parameters using MCMC iterations.

Second, we consider separate modeling where in the first part we use linear mixed models in equations (4.6), and estimate the model parameters. Then, we use the estimated longitudinal outcomes as covariates for modeling the event-time using equation (4.7), and estimate the correspond model parameters. We refer to this model as Model 2.

Finally, we use the proposed latent-class model, and estimate the model parameters in a joint modeling framework using MCMC iterations. This is referred to as

Model 3.

We fit all these three competing models to the simulated dataset, and then compare their performances in terms of the goodness of fit and predictive power. We compute average BIC values and the average mean squared error (AMSE) for three biomarkers based on 100 replicates of the dataset as goodness of fit measures. For Model 2 (separate modeling), we compute BIC from two models separately and add them as the combined BIC value. For assessing the predictive power (corresponding to the three biomarkers), we compute the log pseudo marginal likelihood (LPML) following Gelfand and Dey (1994) [33]. In Table 4.6, we summarize the results. We notice that Model 2 provides the largest BIC and AMSE values, and smallest LPML value. On the other hand, Model 3 results in the smallest BIC and AMSE values, and the largest LPML value. This illustrates that for datasets with a number of subgroups the proposed model provides a more powerful inference.

For the joint modeling of longitudinal traits and event-time the discriminative capability of a model is also evaluated. We use the area under the receiver operating characteristic curve (AUC) for such comparison (Kundu et al., 2023 [52]). The AUC measures how efficiently a joint model can discriminate the subjects with a relapse from the subjects with no relapse (Rizopoulos, 2016 [78]). Let $\pi_i(t+\Delta t|t)$ be the probability that for the i-th subject there is no relapse up to time $t+\Delta t$ given that it is event-free (no relapse) until time t. For any pair of subjects $[i,j]$ who are event-free until time t, the discriminative power of a model is assessed by computing AUC as below:

$AUC = P\left[\pi_i(t+\Delta t|t) < \pi_j(t+\Delta t|t) | (T_i \in (t, t+\Delta t]) \cap (T_j > t+\Delta t)\right]$, where T_i and T_j, respectively, denote the actual event-time for the i-th and the j-th subject. This means that for a fixed time-interval $(t, t+\Delta t)$ if a relapse occurs for the i-th subject but the j-th subject is event-free up to time $t+\Delta t$, then the model must assign a higher non-relapse probability to the j-th subject. We use this criterion to compare different models.

In Table 4.7, we show the AUC values for Model 1 and Model 3 for different choices of Δt, with t=10. We note that the AUC values are always higher for Model 3 than Model 1. This reflects the fact that the proposed latent-class joint model can better discriminate the patients with relapse when there are several sub-populations in the observed dataset. Thus, our simulation studies, in general, establish the usefulness of the proposed model in practice.

TABLE 4.6: BIC, Average MSE and LPML values for the three competing models in the simulation study.

	Model 1	Model 2	Model 3
BIC	248.93	321.51	203.26
AMSE	169.57	223.84	116.39
LPML	286.51	228.33	342.43

TABLE 4.7: AUC values (for t=10, and $\Delta t = 5, 10, 15$) are given for the two competing joint models in the simulation study.

	Model 1	Model 3
AUC(t=10, Δt=5)	0.54	0.73
AUC(t=10, Δt=10)	0.60	0.69
AUC(t=10, Δt=15)	0.58	0.71

4.6 Summary

In this chapter, we develop a Bayesian latent class model for the joint analysis of multivariate longitudinal and event-time data. Traditionally, the fixed covariates determine the latent classes, but we use the longitudinal trajectories (which are indeed influenced by the fixed covariates) of the most important longitudinal outcomes for finding the latent classes. The association between lymphocyte count and the other two biomarkers (neutrophil count and platelet count) can be assessed by looking at the class-specific trajectories of these biomarkers. Our simulation studies also illustrate the usefulness of the proposed modeling approach in the presence of certain sub-groups with different evolution of the biomarkers.

We note that while there is a rich literature on Bayesian latent class modeling, our proposed approach is quite innovative. The existing approaches mostly cluster multiple outcomes altogether, and therefore the interpretation of the clusters is less obvious. Also such approaches are computationally demanding, and sometimes suffer from label switching problem severely. We model multiple outcomes using Bayesian hierarchical models but clustering is done with respect to one of these. The trajectories of the other outcomes are assessed on these latent clusters to find the association among the outcomes at different cluster levels. While this is computationally faster, it also provides a clear understanding of the clusters. Our work, thus, is different from most of the other clustering techniques proposed in the literature except the one proposed in Putter et al. (2008).

There are several limitations of the proposed modeling approach. Our model fails to handle missingness in the longitudinal biomarkers since the missing values will affect the class-membership probabilities significantly. Developing latent class models which can handle missing values in the longitudinal biomarkers and can impute them simultaneously, can be a good research avenue to be explored. Additionally, the latent classes can share some information in terms of the model parameters, and by considering a Dirichlet Process prior (or some other variants of it) we can measure such similarity among different classes. We leave this as an interesting future work.

Chapter 5

Summary and Future Works

5.1 Joint Modeing with ALL dataset

In a longitudinal study where a group of individuals are followed for a certain period of time for the occurrence of one or more events of interest, a joint analysis is highly recommended. In a joint analysis we first model the progression of one or more longitudinal outcomes, and then assess the effects of these outcomes on the event-time through some popular Statistical models, e.g. Cox PH model, Accelerated Failure Time model etc. The merits of a joint analysis of longitudinal outcomes and event-time data over the separate modeling have already been established and discussed in a series of papers (Henderson et al., 2000 [40]; Wang and Taylor, 2001 [96]; Brown et al., 2005 [13]; Chi and Ibrahim, 2006 [15]; Rizopoulos and Ghosh, 2011 [79]; Das, 2016 [19]; Rizopoulos, 2017 [80], and the references therein).

Our work presented in this thesis is motivated by the clinical study mentioned in Section 1.4.1. We note that although the survival rate for ALL is quite high in the developed countries, it is still quite unsatisfactorily low in most of the Asian and the African countries. Since the treatment phase is quite long (nearly two years), most of the poor families in India cannot afford the treatment cost, and are forced to discontinue the treatment. This is the major reason for which the true survival rate for ALL could not be estimated properly in India. The study conducted by TTCRC considered the patients who could complete the treatment, and could also be followed for the next three years. Since a relapse is an indication of the failure of the treatment for a particular patient, we focus on modeling the time to relapse in our analysis. Our analysis provides several interesting insights and meaningful results which can be further investigated for a better healthcare policy in India.

5.2 Contribution of this thesis

In Chapter 2 of this thesis, we develop a Bayesian model for the joint analysis of three outcomes i.e. WBC, ANC and platelet count; and the relapse-time. Multivariate linear mixed models with the Gaussian random effects are used for modeling the

evolution of the biomarkers. A Cox PH model which considers the expected outcomes conditional on the random effects as covariates is used for modeling the relapse-time. For estimating the model parameters in a Bayesian framework, we need to sample from the joint posterior distribution, and we use MCMC for this purpose. Our proposed model can simultaneously impute the missing biomarker values within each MCMC iteration, and can provide the patient-specific dynamic prediction of the non-relapse probability during the treatment and in the follow-up period. We use the posterior predictive distribution for assessing the effects of different doses of two medicines (i.e. 6MP and MTx) on the evolution of the biomarkers and on the hazard rate. Clinically, this analysis recommends a lower dose of 6MP and a higher dose of MTx for a better non-relapse probability. In addition, it shows that the patients classified as "high risk" in the beginning generally experience a lower relapse rate in the follow-up period. Our findings are also validated by extensive simulation studies.

Chapter 3 is focused on developing a Bayesian joint model at the quantile levels since the quantile-based analysis typically provides robust inference. In this chapter, we consider the lymphocyte count as one of the biomarkers, and then observe that the joint distribution of the three biomarkers deviate from a multivariate normal distribution. In addition, the contour plot also indicates that a quantile level analysis is indeed appropriate for the dataset in hand. We develop Bayesian quantile mixed models for the biomarkers, and consider the Brownian motion random effects for better flexibility. We adopt the Bayesian quantile regression approach proposed in Geraci and Bottai (2007) [35], and consider an Asymmetric Laplace Distribution (ALD) for the random errors in the linear quantile mixed models. For the computational ease, we exploit the mixture representation of ALD following Kozumi and Kobayashi (2011) [50], and develop a computationally efficient Gibbs Sampler algorithm for estimating the model parameters. This analysis reflects that 6MP helps to reduce the lymphocyte count, and MTx helps to increase the neutrophil count across all quantiles. However, their effects on platelet count differ from one quantile level to the other. Based on the estimated median non-relapse probability curves we conclude that the patients with a higher lymphocyte count and a lower neutrophil and platelet count experience more relapse than the patients belonging to the higher quantiles for all three biomarkers. This indicates that a higher lymphocyte count itself is not completely responsible for a faster relapse.

In Chapter 4, we consider a different approach, i.e. a latent class Bayesian analysis of the longitudinal outcomes and relapse-time. Here, our goal is to identify the latent classes based on the longitudinal trajectory of the key biomarker, i.e. lymphocyte count. In other words, we distinguish the patients with different evolution of the lymphocyte count. The trajectories for the neutrophil count and the platelet count for these subgroups establish the association among the biomarkers. Based on a joint model, we assess the class-specific effects of the covariates, and estimate the class-specific median non-relapse probability curves. For our dataset the proposed

approach detects two latent classes with distinct features in terms of the biomarkers and their effects on the relapse-time. Practical usefulness of this approach is further investigated through extensive simulation studies.

5.3 Limitations and Future Works

Remembering the well-known comment of Prof. G. Box, "All models are wrong, but some are useful", we mention some limitations of our work presented in this thesis. First of all, progression of the biomarkers and the relapse-time of ALL are greatly influenced by the genetic structure of the patients (see, Moriyama, Relling and Yang, 2015 [58]; Yakota and Kanakura, 2016 [102], and the references therein). Our analysis completely ignores the genetic effects, and thus provides an incomplete inference. However, this is due to the fact that the dataset in hand did not include the genetic information of the children who participated in this study. Typically, huge amount of genetic information is collected for each patient. Statistically, this brings a high-dimensional covariates in the model ($p > n$). However, the models proposed in Chapters 2, 3 and 4, can handle such setting by considering some shrinkage priors (Lasso prior, local-global shrinkage prior, slab and spike prior etc.) for the regression coefficients.

Second, for the quantile level analysis proposed in Chapter 3, it might be of interest to measure the similarity of different quantile levels in terms of the estimated model parameters. A non-parametric Bayesian approach could be used for such inference considering different quantiles as distinct groups. Dunson et al. (2008) [27] proposed Matrix Stick-Breaking Process (MSBP) prior for assessing such similarity. Gaskins and Daniels (2013) [32], Das and Daniels (2014) [20] developed similar priors for simultaneous estimation of the longitudinal outcomes. More recently, Das et al. (2021) [21] developed a dynamic hierarchical Bayesian approach for modeling multiple time-varying groups. This approach can measure the similarity across different groups where the size and the composition of the groups change with time. Similar prior distribution can be used for our Bayesian quantile joint modeling for more interesting results. However, the computational cost could be an issue.

Next, for the Bayesian latent class model proposed in Chapter 4 one can reduce the number of distinct parameters to be estimated by considering an automated clustering of the model parameters. Dirichlet Process (DP) priors are typically used for shrinking the model parameters to a common value (Ferguson, 1973 [29]; Dunson, 2006 [26]; Blei and Jordan, 2006 [11]; Jensen and Shore, 2011 [45]). The stick-breaking representation of DP, proposed in Sethuraman (1994) [85] is used for developing computationally efficient MCMC algorithms (Das et al., 2021 [21]) which can handle complex models. A non-parametric Bayesian treatment of the proposed joint models is definitely an exciting extension of our work.

Finally, we must admit that the dataset provided by TTCRC is indeed an asset, and similar studies are highly recommended for a better understanding of the progression of ALL. An analysis based only on 236 patients is definitely not adequate for answering all clinical questions related to ALL, but similar studies conducted in a larger scale will deepen our understanding on the effectiveness of the maintenance therapy routinely used in India. Such studies will result in similar analyses presented in this thesis, and that in one hand will help the medical experts to determine the optimal drug doses; and on the other hand it will help the policy makers to implement certain change in the existing healthcare policy. We must conclude by noting that even if ALL cannot be cured for all the patients, but the children can definitely live longer with the advanced chemotherapy. Therefore, it is extremely important to conduct similar studies and perform similar analyses presented in this thesis.

BIBLIOGRAPHY

1. Abdelmabood, S., Fouda, A. E., Boujettif, F. & Mansour, A. Treatment outcomes of children with acute lymphoblastic leukemia in a middle-income developing country: high mortalities, early relapses, and poor survival. *Jornal de Pediatria (Versão em Português)* **96,** 108–116 (2020).
2. Adar, S. D. *et al.* Longitudinal analysis of long-term air pollution levels and blood pressure: A cautionary tale from the multi-ethnic study of atherosclerosis. *Environmental health perspectives* **126,** 107001–107003 (2018).
3. Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics* **10,** 1100–1120 (1982).
4. Arora, R. S. & Arora, B. Acute leukemia in children: A review of the current Indian data. *South Asian Journal of Cancer* **5,** 155 (2016).
5. Balan, T. A. & Putter, H. A tutorial on frailty models. *Statistical Methods in Medical Research* **29,** 3424–3454 (2020).
6. Bandyopadhyay, D., Lachos, V. H., Abanto-Valle, C. A. & Ghosh, P. Linear Mixed Models for Skew-Normal/Independent bivariate responses with an application to Periodontal Disease. *Statistics in Medicine* **29,** 2643–2655. (2010).
7. Bassi, F. Longitudinal models for dynamic segmentation in financial markets. *International Journal of Bank Marketing* **35,** 431–446 (2017).
8. Belson, M., Kingsley, B. & Holmes, A. Risk factors for acute leukemia in children: a review. *Environmental health perspectives* **115,** 138–145 (2007).
9. Biswas, J. & Das, K. A Bayesian quantile regression approach to multivariate semi-continuous longitudinal data. *Computational Statistics* **36,** 241–260 (2021).
10. Biswas, J., Ghosh, P. & Das, K. A semi-parametric quantile regression approach to zero-inflated and incomplete longitudinal outcomes. *AStA Advances in Statistical Analysis* **104,** 261–283 (2020).
11. Blei, D. M. & Jordan, M. I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1,** 121–143 (2006).
12. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. **7,** 434–455 (1998).
13. Brown, E. R., Ibrahim, J. G. & DeGruttola, V. A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics* **61,** 64–73 (2005).

14. Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. Deviance information criteria for missing data models. *Bayesian Analysis* **1,** 651–673 (2006).

15. Chi, Y. Y. & Ibrahim, J. Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62,** 432–445 (2006).

16. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34,** 187–202 (1972).

17. Daniels, M. J. & Hogan, J. W. *Missing data in longitudinal studies: Strategies for Bayesian Modeling and Sensitivity Analysis* (Chapman and Hall (CRC Press), 1993).

18. Daniels, M. J. & Pourahmadi, M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89,** 553–566 (2002).

19. Das, K. A semiparametric Bayesian approach for joint modeling of longitudinal trait and event time. *Journal of Applied Statistics* **43,** 2850–2865 (2016).

20. Das, K. & Daniels, M. J. A Semi-parametric Approach to Simultaneous Covariance Estimation for Bivariate Sparse Longitudinal Data. *Biometrics* **70,** 33–43 (2014).

21. Das, K., Ghosh, P. & Daniels, M. J. Modeling multiple time-varying related groups: a dynamic hierarchical Bayesian approach with an application to the health and retirement study. *Journal of the American Statistical Association* **116,** 558–568 (2021).

22. Das, K., Li, J., Fu, G., Wang, Z. & Wu, R. Genome-wide association studies for bivariate sparse longitudinal data. *Human Heredity* **72,** 110–120 (2011).

23. Das, K., Li, R., Huang, Z., Gai, J. & Wu, R. A Bayesian framework for functional mapping through joint modeling of longitudinal and time-to-event data. *International journal of plant genomics* **2012.** https://doi.org/10.1155/2012/680634 (2012).

24. Das, K. *et al.* Dynamic semi-parametric Bayesian models for genetic mapping of complex traits with irregular longitudinal data. *Statistics in Medicine* **32,** 509–523 (2013).

25. Devarajan, K. & Ebrahimi, N. A semi-parametric generalization of the Cox proportional hazards regression model: Inference and applications. *Computational statistics & data analysis* **55,** 667–676 (2011).

26. Dunson, D. B. Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7,** 551–568 (2006).

27. Dunson, D. B., Xue, Y. & Carin, L. The matrix stick-breaking process: flexible Bayes meta-analysis. *Journal of the American Statistical Association* **103,** 317–327 (2008).

28. Farcomeni, A. & Viviani, S. Longitudinal quantile regression in the presence of informative dropout through longitudinal–survival joint modelling. *Statistics in Medicine* **34,** 1199–1213 (2015).

29. Ferguson, T. S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1,** 209–230 (1973).

30. Fieuws, S. & Verbeke, G. Joint modelling of multivariate longitudinal profiles: Pitfalls of the random-effect approach. *Statistics in Medicine* **23,** 3093–3104 (2004).

31. Gabrio, A., Hunter, R., Mason, A. J. & Baio, G. Joint longitudinal models for dealing with missing at random data in trial-based economic evaluations. *Value in Health* **24,** 699–706 (2021).

32. Gaskins, J. T. & Daniels, M. J. A nonparametric prior for simultaneous covariance estimation. *Biometrika* **100,** 125–138 (2013).

33. Gelfand, A. E. & Dey, D. K. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)* **56,** 501–514 (1994).

34. Gelman, A., Mechelen, I. V., Verbeke, G., Heitjan, D. & Meulders, M. Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* **61,** 74–85 (2005).

35. Geraci, M. & Bottai, M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8,** 140–154 (2007).

36. Ghosh, P. & Hanson, T. A. Semiparametric Bayesian approach to multivariate longitudinal data. *Australian and New Zealand Journal of Statistics* **52,** 275–288 (2010).

37. Gomez, G., Julià, O., Utzet, F. & Moeschberger, M. L. Survival analysis for left censored data. *Survival analysis: State of the art* **211,** 269–288 (1992).

38. Gould, A. L. *et al.* Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* **34,** 2181–195 (2015).

39. Guo, X. & Carlin, B. P. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *The American Statistician* **58,** 16–24 (2004).

40. Henderson, R., Diggle, P. & Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480 (2000).

41. Hogan, J. W. & Laird, N. M. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16,** 239–h257 (1997).

42. Hu, W., Li, G. & Li, N. A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine* **28,** 1601–1619 (2009).

43. Huang, X., Li, G., Elashoff, R. M. & Pan, J. A general joint model for longitudinal measurements and competing risks survival data with heterogeneous random effects. *Lifetime Data Analysis* **17,** 80–100 (2011).

44. Ibrahim, J. G. & Molenberghs, G. Missing data methods in longitudinal studies: a review. *Test* **18,** 1–43 (2009).

45. Jensen, S. T. & Shore, S. H. Semiparametric Bayesian modeling of income volatility heterogeneity. *Journal of the American Statistical Association* **106,** 1280–1290 (2011).

46. Kalbfleisch, J. & Prentice, R. *The Statistical Analysis of Failure Time Data. sl: Wiley Interscience* 2nd ed. (2002).

47. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53,** 457–481 (1958).

48. Koenker, R. Quantile regression for longitudinal data. *Journal of Multivariate Analysis* **91,** 74–89 (2004).

49. Koenker, R. & Bassett Jr, G. Regression quantiles. *Econometrica: Journal of the Econometric Society* **46,** 33–50 (1978).

50. Kozumi, H. & Kobayashi, G. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* **81,** 1565–1578 (2011).

51. Kulkarni, H., Biswas, J. & Das, K. A joint quantile regression model for multiple longitudinal outcomes. *Advances in Statistical Analysis* **103,** 453–473 (2019).

52. Kundu, D., Sarkar, P., Gogoi, M. & Das, K. A Bayesian joint model for multivariate longitudinal and time-to-event data with application to ALL maintenance studies. *Journal of Biopharmaceutical Statistics*. https://doi.org/10.1080/10543406.2023.2187413 (2023).

53. Lagakos, S. W. General right censoring and its impact on the analysis of survival data. *Biometrics* **35,** 139–156 (1979).

54. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38,** 963–974 (1982).

55. Lin, H., Turnbull, B. W., McCulloch, C. E. & Slate, E. H. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97,** 53–65 (2002).

56. Liu, L., Huang, X. & O'Quigley., J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics* **64,** 950–958 (2008).

57. Liu, Y., Liu, L. & Zhou, J. Joint latent class model of survival and longitudinal data: An application to CPCRA study. *Computational Statistics & Data Analysis* **91,** 40–50 (2015).

58. Moriyama, T., Relling, M. V. & Yang, J. J. Inherited genetic variation in childhood acute lymphoblastic leukemia. *Blood, The Journal of the American Society of Hematology* **125,** 3988–3995 (2015).

59. Moundele, C. M., Odongo, L. & Banzouzi, B. N. M. K. An Estimator of Baseline Intensity by the Method of Wavelets. *Applied Mathematical Sciences* **13,** 547–558 (2019).

60. Mustefa, Y. A. & Chen, D. G. Accelerated failure-time model with weighted least-squares estimation: application on survival of HIV positives. *Archives of Public Health* **79.** https://doi.org/10.1186/s13690-021-00617-0 (2021).

61. Muthén, B. & Shedden, K. Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55,** 463–469 (1999).

62. Pan, J. & Mackenzie, G. On modelling mean-covariance structures in longitudinal studies. *Biometrika* **90,** 239–244 (2003).

63. Papageorgiou, G., M. M. Mokhles, J. J. T. & Rizopoulos., D. Individualized dynamic prediction of survival under time-varying treatment strategies. *arXiv preprint arXiv:1804.02334* (2018).

64. Papageorgiou, G., Mauff, K., Tomer, A. & Rizopoulos, D. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual Review of Statistics and Its Application* **6,** 223–240 (2019).

65. Picchini, U., Gaetano, A. D. & Ditlevsen, S. Stochastic differential mixed-effects models. *Scandinavian Journal of Statistics* **37,** 67–90 (2010).

66. Pourahmadi, M. Joint mean-covariance model with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86,** 677–690 (1999).

67. Pourahmadi, M. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika* **87,** 425–435 (2000).

68. Prentice, R. L. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342 (1982).

69. Proust-Lima, C., M. Séne, J. M. T. & Jacqmin-Gadda, H. Joint latent class models for longitudinal and time-to-event data: a review. *Statistical Methods in Medical Research* **23,** 74–90 (2014).

70. Proust-Lima, C. & Taylor, J. M. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics* **10,** 535–549 (2009).

71. Pui, C. H. & Evans, W. E. A 50-year journey to cure childhood acute lymphoblastic leukemia. *Seminars in Hematology* **50,** 185–196 (2013).

72. Pui, C. H., Yang, J. J., Bhakta, N. & Rodriguez-Galindo, C. Global efforts toward the cure of childhood acute lymphoblastic leukaemia. *The Lancet Child and Adolescent Health* **2,** 440–454 (2018).

73. Putter, H., Vos, T., de Haes, H. & van Houwelingen, H. Joint analysis of multiple longitudinal outcomes: application of a latent class model. *Statistics in Medicine* **27,** 6228–6249 (2008).

74. Renbarger, J. L., McCammack, K. C., Rouse, C. E. & Hall, S. D. Effect of race on vincristine-associated neurotoxicity in pediatric acute lymphoblastic leukemia patients. *Pediatric Blood and Cancer* **50,** 769–771 (2008).

75. Rhein, P. *et al.* Intermediate-Risk Acute Lymphoblastic Leukemia (ALL) Patients with and without Relapse Differentially Depend on Survival Signals From Microenvironment. *Blood* **118,** 752 (2011).

76. Rizopoulos, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67,** 819–829 (2011).

77. Rizopoulos, D. *Joint models for longitudinal and time-to-event data: With applications in R.* (CRC press, 2012).

78. Rizopoulos, D. The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *Journal of Statistical Software* **72,** 1–45 (2016).

79. Rizopoulos, D. & Ghosh, P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30,** 1366–1380 (2011).

80. Rizopoulos, D., Molenberghs, G. & Lesaffre, E. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* **59,** 1261–1276 (2017).

81. Rodrigues, A. S., Calsavara, V. F., Silva, F. I., Alves, F. A. & Vivas, A. P. Use of interval-censored survival data as an alternative to Kaplan-Meier survival curves: studies of oral lesion occurrence in liver transplants and cancer recurrence. *Applied Cancer Research* **38,** 1–10 (2018).

82. Rubin, D. B. Inference and missing data. *Biometrika* **63,** 581–592 (1976).

83. Sahu, S. K., Dey, D. K., Aslanidou, H. & Sinha, D. A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis* **3,** 123–137 (1997).

84. Schafer, J. L. & Yucel, R. M. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* **11,** 437–457 (2002).

85. Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica* **4,** 639–650 (1994).

86. Sithole, J. S. & Jones, P. W. Bivariate Longitudinal Model for Detecting Prescribing Change in Two Drugs Simultaneously with Correlated Errors. *Journal of Applied Statistics* **34,** 339–352 (2007).

87. Sy, J. P., Taylor, J. M. G. & Cumberland, W. G. A Stochastic model for the analysis of bivariate longitudinal AIDS data. *Biometrics* **53,** 542–555 (1997).

88. Thiebaut, R., Jacqmin-Gadda, H., Chene, G., Leport, C. & Commenges, D. Bivariate linear mixed models using SAS PROC MIXED. *Computer Methods and Programs in Biomedicine* **69,** 249–256 (2002).

89. Tian, L., Zucker, D. & Wei, L. J. On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association* **100,** 172–183 (2005).

90. Tomer, A., Nieboer, D., Roobol, M. J., Steyerberg, E. W. & Rizopoulos., D. Personalized schedules for surveillance of low-risk prostate cancer patients. *Biometrics* **75,** 153–162 (2019).

91. Tsiatis, A. A. & Davidian, M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14,** 809–834 (2004).

92. Varghese, B., Joobomary, A. A. & Savida, P. Five-Year survival rate and the factors for risk-directed therapy in acute lymphoblastic leukemia. *Indian Journal of Medical and Paediatric Oncology* **39,** 301 (2018).

93. Verweij, P. J. & van Houwelingen, H. C. Time-dependent effects of fixed covariates in Cox regression. *Biometrics* **51,** 1550–1556 (1995).

94. Wang, C. & Hall, C. B. Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine* **29,** 671–679 (2010).

95. Wang, H., Snapp, S. S., Fisher, M. & Viens, F. A Bayesian analysis of longitudinal farm surveys in Central Malawi reveals yield determinants and site-specific management strategies. *Plos one* **14,** 1–7 (2019).

96. Wang, Y. & Taylor, J. M. G. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96,** 895–905 (2001).

97. Williamson, P. R., Kolamunnage-Dona, R., Philipson, P. & Marson, A. G. Joint modelling of longitudinal and competing risks data. *Statistics in Medicine* **27,** 6426–6438 (2008).

98. Wong, K. Y., Zeng, D. & Lin, D. Y. Semiparametric latent-class models for multivariate longitudinal and survival data. *Annals of Statistics* **50,** 487–510 (2022).

99. Wu, W. B. & Pourahmadi, M. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90,** 831–844 (2003).

100. Xu, J. & Zeger, S. L. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50,** 375–387 (2001).

101. Yang, M., Luo, S. & DeSantis, S. Bayesian quantile regression joint models: inference and dynamic predictions. *Statistical Methods in Medical Research* **28,** 2524–2537 (2019).

102. Yokota, T. & Kanakura, Y. Genetic abnormalities associated with acute lymphoblastic leukemia. *Cancer Science* **107,** 721–725 (2016).

103. Yu, K. & Moyeed, R. A. Bayesian quantile regression. *Statistics & Probability Letters* **54,** 437–447 (2001).

104. Zavrakidis, I., Jóźwiak, K. & Hauptmann, M. Statistical analysis of longitudinal data on tumour growth in mice experiments. *Scientific Reports* **10,** 1–11 (2020).

105. Zeng, D. & Cai, J. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *The Annals of Statistics* **33,** 2132–2163 (2005).

106. Zeng, D. & Lin., D. Y. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association* **102,** 1387–1396 (2007).

107. Zhang, H. & Huang, Y. Quantile regression-based Bayesian joint modeling analysis of longitudinal–survival data, with application to an AIDS cohort study. *Lifetime Data Analysis* **26,** 339–368 (2020).

108. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Annals of Translational Medicine* **6.** https://doi.org/10.21037/atm.2018.02.12 (2018).

109. Zhu, L. *et al.* Semiparametric transformation models for joint analysis of multivariate recurrent and terminal events. *Statistics in Medicine* **30,** 3010–3023 (2011).

Biography of the Author

Damitri Kundu was born on March 15, 1994. She received her B.Sc (Statistics Hons) from St. Xaviers College Kolkata and M.Stat from Indian Statistical Institute, Kolkata in 2015 and 2017 respectively. She started her career as an Associate Data Scientist in an MNC in Bangalore. Later, she joined as a Junior Research Fellow in Statistics at Indian Statistical Institute, Kolkata in June, 2018 where she is currently a PhD student in the Applied Statistical Unit.

Her research interests include variable selection, longitudinal data analysis, latent class modeling, multivariate quantile contours, etc . She has published four journal articles with three more on the way.

Apart from research work, she is loves to solve puzzles and is a coding enthusiast. In here free time she enjoys listening to music, drawing and watching anime.