Statistical modelling of time to relapse or onset of metastatic disease following treatment for cancer: relationship to residual/resistant disease and re-growth

Crispin Musicha

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Leeds Institute of Clinical Trials Research

School of Medicine

September, 2019

This candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to thank my supervisors, Professor Linda Sharples and Associate Professor David Cairns, for their tireless support throughout the years I have been working on this project. I want to thank Linda for having been there from the very beginning when she advertised this PhD, to me moving from Malawi to Leeds, and still continuing to be ever present despite moving to London in 2017 with regular supervision meetings each and every week regardless of her busy schedule. I will also remember the crucial role she played in the first year when the project seemed to be going nowhere before she introduced me to structural equation models which I had never used before. I thoroughly enjoyed learning from Linda.

I also thank David, who despite joining the project a year on, brought with him a calming and reassuring presence when things seemed to be getting nowhere. He also helped me understand more about Myeloma and CLL, beyond just fitting survival models to variables I otherwise would have no understanding of. David was always happy to help whenever needed and to deal with the day to day issues.

I would also like to thank retired Professor Walter Gregory, who introduced me to his passion, the modelling of residual disease and tumour growth rate in cancer. I would also like to thank Professor Julia Brown for providing with all the necessary support including office space at the CTRU.

Finally, I would like to thank my wife Yamikani, for always being there for me and the children as I struggled with MCMC.

Abstract

Response to treatment for some cancers may result in a proportion of the patients being permanently cured of the disease, while the rest remain at risk of a relapse or progression. Standard survival analysis approaches assume that all patients will eventually experience the event of interest and so are not appropriate in these situations. The residual disease and tumour re-growth rate following treatment are two of the important predictors of outcomes among cancer patients following treatment. However, in most cases, data on these measures is not always available.

Using data from multiple myeloma and chronic lymphocytic leukaemia trials conducted at the Leeds Institute of Clinical Trials Research, we investigated the role of the residual disease and other important factors associated with time to relapse and overall survival, as well as the probability of being cured of the malignancy following treatment. As the multiple myeloma trial also collected data on bio-markers for tumour growth rate, we used the structural equation modelling approach to investigate the role of growth rate on the survival outcomes using available statistical software. We then extended these models to also investigate the association of growth and the probability of being cured after treatment in a Bayesian framework including in situations where not all patients will relapse or die after treatment.

This work demonstrates that it is feasible to use structural equation modelling within survival analysis to investigate the role of latent variables on time to event outcomes even in situations where not all patients will relapse or die after treatment.

Contents

1 Background **3**

1.1 Aim 9

1.2 Outline of thesis 11

2 Methods for analysing time-to-event data **12**

2.1 Introduction 12

2.2 Overview of TTE data 13

2.3 Non-parametric approaches 15

2.4 Semi-parametric approaches 16

2.4.1 The Cox PH model 16

2.4.2 Partial likelihood for the Cox PH model 17

2.5 Parametric models 18

2.6 Cure rate models for the proportion that will never relapse following treatment 20

2.7 Cure rate models for estimating the proportion with OS similar to the general population 21

2.7.1 Calculating the population hazard and survival rate 22

2.8 Bayesian TTE models 23

2.9 Overview of parameter estimation in TTE models 23

2.9.1 ML for fitting standard models 24

2.9.2 Using the EM algorithm 25
2.9.3 Using Markov Chain Monte Carlo methods to fit Bayesian models 27
2.10 Assessing the model fit in TTE models 32
2.10.1 Schoenfeld residuals 32
2.10.2 Other residuals and methods for assessing model fit 33
2.10.3 Assessing model fit in cure rate models 34
2.10.4 Checking model convergence in Bayesian models 34
2.11 Model selection 36
2.11.1 Choosing between models using the Deviance Information Criterion and other ways of model selection 36
2.12 Summary 38

3 Exploring factors affecting survival outcomes in patients recruited in UK Myeloma and CLL trials 40

3.1 Introduction 40
3.1.1 Outcomes and variable descriptions 41
3.2 The Myeloma data set 42
3.2.1 Predictors of TTE outcomes in Myeloma 43
3.2.2 Descriptive summary of the categorical variables in Myeloma . . . 45
3.2.3 Overall TTR and OS patterns in Myeloma 47
3.2.4 Comparing TTR patterns by categorical variables in Myeloma . . . 47
3.2.5 Comparing observed OS patterns by categorical variables in Myeloma 50
3.2.6 Descriptive summaries of the continuous variables in Myeloma . . 52
3.2.7 Modelling independent factors associated with TTR in Myeloma . 53
3.2.8 Modelling independent factors associated with OS in Myeloma . . 56
3.2.9 Multivariable models for TTR and OS in Myeloma 60
3.2.10 Sensitivity analysis 62

3.3 Application of standard survival methods to CLL data 66
3.3.1 Predictors of TTE outcomes in CLL 66
3.3.2 Descriptive summary of the categorical variables in CLL 66
3.3.3 Summary of the continuous variables by TTR and OS 69
3.3.4 Overall TTR and OS patterns in CLL 70
3.3.5 Comparing TTR patterns by categorical variables in CLL 71
3.3.6 Comparing OS patterns by categorical variables in CLL 71
3.3.7 Modelling independent factors associated with TTR in CLL 71
3.3.8 Modelling independent factors associated with OS in CLL 76
3.3.9 Multivariable models for TTR and OS in CLL 79
3.4 Handling missing data 83
3.4.1 Missing data in the Myeloma dataset 85
3.4.2 Baseline characteristics of participants with complete and incomplete data in the Myeloma dataset 85
3.4.3 Univariable Cox PH models for TTR and OS in Myeloma fitted to imputed data 88
3.4.4 Multivariable Cox PH models for TTR and OS in Myeloma fitted to imputed data 91
3.4.5 Increasing the number of imputations versus using Bayesian models to account for missing data in univariable Cox PH models for TTR and OS in Myeloma 94
3.4.6 Missing data in the CLL dataset 98
3.4.7 Univariable Cox PH models for TTR and OS in CLL fitted to imputed data 100
3.4.8 Multivariable Cox PH models for TTR in CLL fitted to imputed data 103
3.5 Summary 104

4 Estimating the cured proportion following treatment in cancer TTE models 106

4.1 Introduction 106

4.2 Parametric mixture models 107

4.3 Parametric non-mixture models 109

4.4 Semi-parametric mixture cure models 111

4.5 Cure rate models that incorporate population survival data 112

4.5.1 Population mixture models 112

4.5.2 Population non-mixture models 113

4.6 Application of cure rate models to the Myeloma and CLL datasets 114

4.6.1 Prior distributions for model parameters in Bayesian cure rate models 115

4.7 Cure rate models without covariates in Myeloma 115

4.7.1 Cure rate models with covariates in Myeloma 118

4.8 Estimating the proportion whose OS returns to that of general population in Myeloma 122

4.8.1 Modelling covariates in cure rate models for OS in Myeloma 124

4.9 Missing data in cure rate models applied to Myeloma 128

4.10 Multivariable cure rate models in Myeloma 130

4.11 Comparing estimates in standard and mixture models in Myeloma 133

4.12 Cure rate models applied to the CLL dataset 134

4.12.1 Covariates in cure rate models for OS in CLL 137

4.13 Summary 139

5 Using structural equation modelling to evaluate the effect of a latent covariate on TTE outcomes 149

5.1 Introduction 149

5.1.1 The measurement model 151

5.1.2 Measurement model with binary observed variables 152
5.1.3 The structural model 153
5.2 Estimating parameters in SEMs 153
5.2.1 Maximum likelihood estimation for SEMs 154
5.2.2 Using Bayesian methods to fit SEMs 155
5.2.3 Bayesian estimation of parameters in measurement models 156
5.2.4 TTE model with a latent variable as a predictor 159
5.3 Ensuring model identifiability 161
5.3.1 Identifiability in a measurement model with 3 observed variables . 162
5.4 Applications to simulated TTE data 164
5.4.1 Models with no censoring 166
5.4.2 Models with some censoring 167
5.4.3 Models with more than 30% censoring 169
5.5 Summary 171

6 Using SEMs to model the role of unobserved tumour growth on TTE outcomes: application to Myeloma data 175

6.1 Introduction 175
6.1.1 Brief description of the data for modelling the effect of tumour growth on the TTE outcomes in Myeloma 177
6.2 TTE models with growth and log-RD in Myeloma 179
6.2.1 Measurement model for the growth rate 181
6.2.2 Prior distributions and constraints 182
6.3 The effect of tumour growth on the TTR in Myeloma 184
6.4 The effect of tumour growth and log-RD on the TTR in Myeloma 186
6.5 The effect of tumour growth on OS in Myeloma 188
6.6 The effect of tumour growth and log-RD on OS in Myeloma 190

6.7 Modelling the effect of tumour growth in cure rate models for TTR 193
6.8 Modelling the effect of tumour growth and log-RD in cure rate models for TTR 196
6.9 Tumour growth rate as a covariate in population mixture and PTC models 205
6.10 Tumour growth rate and log-RD as covariates in population mixture and PTC models 209
6.11 Checking for the possibility of a cured proportion 216
6.12 Relaxing the multivariate normal assumption in SEM 219
6.12.1 The effect of growth and log-RD on TTR with paraprotein as binary covariate 221
6.12.2 The effect of growth and log-RD on OS with paraprotein as binary covariate 223
6.12.3 Comparison with SEMs based on continuous variables 224
6.13 Summary 225

7 Discussion **232**
7.1 Introduction 232
7.2 Results in context 235
7.3 Limitations 236
7.4 Further work 237

Appendices **238**

A Cure rate models in OpenBUGS **239**
A.1 Mixture model code in OpenBUGS 239
A.2 Population mixture model code in OpenBUGS 240
A.3 Population PTC model code in OpenBUGS 241

B Simulated TTE model with a latent covariate **243**

C OpenBUGS code for fitting Weibull TTE models with SEM extensions 247

C.1 Bayesian Weibull TTE model with latent tumour growth 247

C.2 Bayesian Cox TTE model with latent tumour growth 249

D OpenBUGS code for fitting the mixture, PTC and SEM models 251

D.1 Bayesian Weibull mixture or PTC model with latent tumour growth 251

D.2 Bayesian Weibull population mixture or PTC models with latent tumour growth 253

E Assessment of model fits in Myeloma using scaled Schoenfeld residuals 256

E.1 Assessing the multivariable Cox PH model for TTR in Myeloma 257

E.2 Assessing the multivariable Cox PH model for OS in Myeloma 258

F Assessment of model fits in CLL using scaled Schoenfeld residuals 259

F.1 Assessing the multivariable Cox PH model for TTR in CLL 259

F.2 Assessing the multivariable Cox PH model for OS in CLL 260

List of Tables

3.1 A summary of the covariates used in the analysis of the Myeloma dataset showing proportion missing for each covariate. The total number of patients in the analysis was 427 46

3.2 Summary of categorical variables in the Myeloma dataset by TTR and OS outcomes. The number of available observations for each variable are shown in parentheses next to the variable name. 48

3.3 Summary by continuous variables in the Myeloma dataset. Age and log-RD, which was already log-transformed, were not z-standardised (*). The summaries for paraprotein, beta2 and albumin were based on 334, 310 and 426 patients respectively. 52

3.4 Log-hazard ratio estimates for each covariate from univariable Cox PH, Exponential and Weibull models for TTR applied to the Myeloma dataset in a CC analysis (details of n for each covariate are shown based on the summary in Table 3.1). Estimates of the intercept for the Exponential and Weibull models as well as the shape parameter (γ) for the Weibull model are also reported. 55

3.5 Log-hazard ratio estimates for each covariate from univariable Cox PH, Exponential and Weibull models for OS applied to the Myeloma dataset in a CC analysis (details of n for each covariate are shown based on the summary in Table 3.1). Estimates of the intercept for the Exponential and Weibull models as well as the shape parameter (γ) for the Weibull model are also reported. 59

3.6 Multivariable Cox PH, Exponential and Weibull models for TTR and OS in the Myeloma dataset. The multivariable TTR and OS models were fitted to data from $n = 125$ patients 63

3.7 Log-hazard ratio estimates from univariable Cox PH models for TTR and OS applied to the same Myeloma dataset used in the final multivariable models with $n = 125$ to enable comparisons with multivariable models . . 64

3.8 Log-HR and adjusted log-HR estimates from univariable and multivariable Cox PH models for TTR and OS applied to the Myeloma dataset ignoring t(4:14) and gain(1q21), $n = 243$. 65

3.9 A summary of the covariates used in the analysis of the CLL dataset showing proportion missing for each covariate. The total number of patients in the analysis was 325 67

3.10 Summary of categorical variables in the CLL dataset by TTR and OS outcomes. For VH mutation*, $n = 286$, while $n = 301$ for p-deletion** . . . 68

3.11 Estimates of log-HRs from univariable Cox PH, Exponential and Weibull models for TTR in the CLL dataset. In the Exponential and Weibull models, estimates of the intercept in the model for the log-hazard are presented on top of the log-HR, while the shape parameter (γ) is also estimated in the Weibull models. There were 286 patients with VH mutation* records and 301 with p-deletion**. For the other covariates, models were fitted with $n = 325$ as they had no missing data. 75

3.12 Estimates of log-HRs from univariable Cox PH, Exponential and Weibull models for OS in the CLL dataset. In the Exponential and Weibull models, estimates of the intercept in the model for the log-hazard are presented on top of the log-HR, while the shape parameter (γ) is also estimated in the Weibull models. There were 286 patients with VH mutation* records and 301 with p-deletion**. For the other covariates, models were fitted with $n = 325$ as they had no missing data. 78

3.13 Estimates of adjusted log-HRs from Cox PH, Exponential and Weibull models for TTR and OS in the CLL dataset. The final TTR model was fitted to data from 286 individuals while the OS model had data from 301 patients 81

3.14 Univariable Cox PH models for TTR and OS in the CLL dataset fitted to data from 286 individuals in the multivariable TTR model and the 301 patients in the final OS model respectively 82

3.15 Missingness patterns in the Myeloma dataset. The (+) represents observed data while the (-) represents missing data. 86

3.16 Comparison of incomplete and completed binary and continuous variables in the Myeloma dataset. Note: SD for albumin not calculated as only one value was missing. n_{cc} represents number of complete cases while n_{ic} is the number of incomplete cases. For the incomplete cases, the means and proportions are averaged over the 10 datasets. 87

3.17 Log-HR estimates and SEs, in brackets, from CC univariable Cox PH models fitted to covariates with missing data in the Myeloma dataset and equivalent results obtained via MI in Stata including the fraction of missing information (FMI) for each covariate. 90

3.18 Adjusted log-HR estimates and SEs, in brackets, from multivariable Cox PH models for TTR and OS in Myeloma fitted to covariates using MI in Stata to account for missing covariate data. The FMI and 95% CIs are also reported. 94

3.19 Log-HR (95% CI) estimates and SEs as well as the FMI from univariable Cox PH models for TTR and OS in Myeloma fitted to covariates using MI in Stata to account for missing covariate data based on $m = 40$ imputations. The estimated log-HRs (95% CrI) and SEs are also reported from Bayesian Cox PH models with imputation models for missing data for each covariate. 97

3.20 Missingness pattern in the CLL dataset. The (+) represents observed data while the (-) represents missing data. 100

3.21 Proportion of those who had hyperdiploidy and p-deletion in CLL among the CCs compared to proportions after MI. n_{cc} represents number of complete cases while n_{ic} is the number of incomplete cases. 100

3.22 Log-HR estimates and SEs, in brackets, from CC univariable Cox PH models fitted to covariates with missing data in the CLL dataset and equivalent results obtained via MI in Stata including the fraction of missing information (FMI) for each covariate. 101

3.23 Adjusted log-HR estimates and SEs, in brackets, from multivariable Cox PH models for TTR in CLL fitted to covariates using MI in Stata to account for missing covariate data. The FMI and 95% CIs are also reported. . . . 103

4.1 Estimates of π, the scale (λ) and shape (γ) parameters from Bayesian Weibull mixture and PTC models for TTR in the Myeloma dataset. 116

4.2 Log-RD ($n = 427$), paraprotein ($n = 334$), beta2 ($n = 310$) and albumin ($n = 426$) as independent covariates in mixture and PTC models for TTR in Myeloma. In the cure models, estimates of the intercept and log-ORs from all models are presented, while estimates from the TTR models include the log-HRs in all models, and log-HRs as well as the intercept and shape parameter (γ) in the Weibull models. 142

4.3 Estimates of π and the scale (λ) and shape (γ) parameters from Weibull population mixture and PTC models for OS in the Myeloma dataset fitted using ML and MCMC (Bayesian) 143

4.4 Log-RD ($n = 427$), paraprotein ($n = 334$), beta2 ($n = 310$), albumin ($n = 426$) as independent covariates in mixture and PTC models for OS in Myeloma. In the cure models, estimates of the intercept and log-ORs from all models are presented, while estimates from the OS models include the log-HRs in all models, and log-HRs as well as the intercept and shape parameter (γ) in the Weibull models. 144

4.5 Estimates of log-OR and log-HRs for paraprotein and beta2 as covariates in Bayesian mixture models for TTR and OS in Myeloma with imputation models to account for missing data. In the cure models, estimates of the intercept and log-ORs are presented, while estimates from the TTE models include log-HRs as well as the intercept and shape parameter (γ) in the Weibull models. 145

4.6 Log-RD, paraprotein, beta2 and albumin as covariates in a multivariable ML mixture model for OS in Myeloma fitted to data from $n = 243$ patients. In the cure model, log-OR estimates as coefficients of each covariate, the SE and 95% CI are reported, while estimates in the OS model represent log-HRs, their SE and 95% CI. 146

4.7 Bayesian mixture and PTC models for π and TTR and OS with a Weibull model in the CLL dataset. Estimates of π, the scale (λ) and the shape (γ) are reported for each model. 147

4.8 Log-RD, stage and sex as covariates in univariable mixture and PTC models for those whose OS returns to that of disease-free individuals in the CLL dataset. In the cure models, estimates of the intercept and log-ORs from both models are presented, while estimates from the OS models include the log-HRs as well as the intercept and shape parameter (γ) in the Weibull models. 148

5.1 Measurable and estimable parameters in a measurement model with 3 observed variables with ϕ fixed at 1. Here, S is the sample covariance matrix, while Σ is the model covariance matrix. 163

5.2 Performance measures for evaluating model fit. J is the number of simulations, $\bar{\hat{\boldsymbol{\theta}}}$ is the average estimate over all simulations, SE($\hat{\boldsymbol{\theta}}_j$) is the SE of the estimate of $\boldsymbol{\theta}$ in the j-th dataset, $j = 1, 2, \ldots, J$, while 3.92 represents twice the z-value (1.96) of the 95% CI for the standard normal distribution. 166

5.3 Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ without censoring. 168

5.4 Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ with 18% censoring on average. 170

5.5 Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ with 39% censoring on average. 172

6.1 Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for TTR fitted to $n = 243$ Myeloma patients with tumour growth rate as a covariate. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 186

6.2 Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for TTR fitted to $n = 243$ Myeloma patients with tumour growth rate adjusted for log-RD as covariates. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) for all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 189

6.3 Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for OS fitted to $n = 243$ Myeloma patients with tumour growth rate as a covariate. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 191

6.4 Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for OS fitted to $n = 243$ Myeloma patients with tumour growth rate adjusted for log-RD as covariates. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) for all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 193

6.5 Bayesian Weibull mixture and PTC models for the proportion that will never relapse and the TTR with growth rate as covariate on both π and λ fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-OR as a coefficient for growth (α_1) are presented. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 197

6.6 Bayesian Weibull mixture and PTC models for the proportion that will never relapse and the TTR with growth rate and log-RD as covariates on both π and λ fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-ORs as coefficients for growth (α_1) and log-RD (α_2) respectively, are presented. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) respectively from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 204

6.7 Bayesian Weibull mixture and PTC models for the proportion with OS similar to general population fitted to Myeloma data with tumour growth rate as covariate on both π and the OS fitted to the Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-OR as a coefficient for growth (α_1) are presented. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 210

6.8 Bayesian Weibull mixture and PTC models for the proportion with OS similar to the general population fitted to Myeloma data with tumour growth rate and log-RD as covariates on both π and the OS fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-ORs as coefficients for growth (α_1) and log-RD (α_2) respectively, are presented. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) respectively from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported 229

6.9 Weibull ML models for the TTR fitted to Myeloma data based on complete cases ($n = 243$) with tumour growth rate only (A) and tumour growth rate and log-RD (B) as covariates on the TTR with paraprotein treated as binary. In the TTR model, the intercept (β_0) and shape parameter (γ), and the log-HR, as a coefficient for growth (β_1) in (A) and for both growth (β_1) and log-RD (β_2), are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 230

6.10 Weibull ML models for the OS fitted to Myeloma data based on complete cases ($n = 243$) with tumour growth rate only (A) and tumour growth rate and log-RD (B) as covariates on the OS with paraprotein treated as binary. In the OS model, the intercept (β_0) and shape parameter (γ), and the log-HR, as a coefficient for growth (β_1) in (A) and for both growth (β_1) and log-RD (β_2), are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported. 231

List of Figures

1.1 General model for time to relapse and subsequently death with both observed covariates and unobserved residual disease and tumour re-growth rate 4

3.1 Derivation of the Myeloma IX final data set 44
3.2 K-M estimates of the overall TTR (left) and OS (right) in the Myeloma dataset. Dotted lines represent 95% CI limits. 49
3.3 K-M estimates of the overall TTR and OS patterns in the Myeloma dataset 50
3.4 K-M estimates of OS by categorical variables in the Myeloma dataset . . 51
3.5 Scaled Schoenfeld residuals plotted against time from the univariable Cox PH models for TTR in the Myeloma dataset ignoring missing data . . 57
3.6 Scaled Schoenfeld residuals plotted against time from the univariable Cox PH models for OS in the Myeloma dataset ignoring missing data . . . 60
3.7 Box plots for log-RD and age at randomisation in the CLL dataset 69
3.8 K-M estimates of the overall TTR (left) and OS (right) in the CLL dataset. Dotted lines represent 95% CI limits. 70
3.9 K-M plots for TTR by categorical variables in the CLL dataset 72
3.10 K-M plots for OS by categorical variables in the CLL dataset 73
3.11 Scaled Schoenfeld residuals plotted against time from univariable Cox PH models for TTR in the CLL dataset ignoring missing data 76

3.12 Scaled Schoenfeld residuals plotted against time from univariable Cox PH models for OS in the CLL dataset ignoring missing data 79
3.13 Histograms of z-standardised paraprotein and beta2 overlaid by imputed values from the first imputation (grey) and boxplots showing the distributions of the observed and imputed z-standardised paraprotein and beta2 from 10 imputed datasets (bottom) 88
3.14 Schoenfeld residuals from univariable Cox PH models for TTR with z-standardised paraprotein, beta2, hyperdiploidy (hyperdip), t(4:14), t(11:14) and gan(1q21) as covariates, plotted against time since RD measurement using the first imputed dataset 91
3.15 Trace and density plots for the coefficient of paraprotein (A) and t(11;14) (B) in Cox PH models for TTR. (C) and (D) are trace and density plots of the coefficients of paraprotein and t(11;14) respectively in Cox PH models for OS. 98
3.16 Schoenfeld residuals from univariable Cox PH models for TTR with VH mutation (vhmut), and p-deletion (pdel) as covariates, plotted against time since RD measurement using the second imputated dataset in CLL . 102

4.1 Model estimated and K-M survivor curves from Bayesian Weibull mixture (A) and PTC (B) models for TTR in the Myeloma dataset 117
4.2 Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian Weibull mixture (A) and PTC (B) models for TTR fitted to the Myeloma dataset. 118

4.3 Model estimated survivor functions by selected log-RD values from the PH mixture model fitted using smcure to the Myeloma data. Graph (A) shows curves with decreasing log-RD while (B) are curves with increasing log-RD. The grey horizontal line shows the estimated π from the Bayesian mixture model while dotted horizontal lines are upper and lower limits of the 95% CrI for the cured proportion. The solid black line represents the K-M estimate of the survivor function for all patients. 122

4.4 Estimated K-M curve and overall survivor (blue dotted line) curves and OS curves for those not cured (red dotted line) from the Bayesian Weibull population mixture (A) and Bayesian population PTC (B) models respectively. 124

4.5 Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian population Weibull mixture (A) and Bayesian Weibull PTC (B) models for OS fitted to the Myeloma dataset. 125

4.6 Trace and density plots from the Bayesian mixture model for OS fitted to the Myeloma dataset with paraprotein as a covariate on both π and λ. The model was fitted to data from 334 patients. 128

4.7 Trace and density plots of the coefficients for paraprotein in the Myeloma dataset on the cured proportion, alpha1 (top) with the effect on the TTR on the left (A) and on OS (B) on the right. Estimates of the effect of paraprotein (beta1) on the TTE (TTR or OS) are shown on the second row in the respective models. 131

4.8 Estimates of the log-HR for log-RD in the TTR standard (top) and cure rate models (bottom). Dots represent point estimates of the log-HR while lines are the 95% CIs in the models fitted using ML or CrIs in the Bayesian models. 134

4.9 Estimates of the log-HR for log-RD in the OS standard (top) and cure rate models (bottom). Dots represent point estimates of the log-HR while lines are the 95% CIs in the models fitted using ML or CrIs in the Bayesian models. 135

4.10 Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian Weibull mixture and PTC models (A and B) for TTR as well as Bayesian Weibull population mixture and PTC models (C and D) for OS fitted to the CLL dataset. 137

5.1 Histograms depicting the distribution of the latent variable as a covariate on the TTE model and in the measurement models from simulated datasets with $n = 500$. The bi-modal distributions illustrate problems with accuracy in some of the simulations. 173

6.1 K-M plots for the TTR and OS for 243 Myeloma patients with paraprotein, beta2 and albumin data. Dotted lines represent 95% CI limits. 178

6.2 Pairs plot showing the distributions of the log-transformed biomarkers as well as the correlation between them 179

6.3 Boxplots showing log-transformed paraprotein, beta2 and albumin by relapse and survival status at the end of the study in Myeloma 180

6.4 SEM model for TTR or OS with tumour growth rate and RD 181

6.5 Trace and density plots of the effect growth on the TTR from the Bayesian Cox PH model (A) and from the Bayesian Weibull model (B) in the Myeloma dataset. 187

6.6 Trace plots for the log-ORs for growth (beta1) and log-RD (beta2) as well as factor loadings tau[1] - tau[3] and variances for paraprotein (var[1]), beta2 (var[2]) and albumin (var[3]) from the Bayesian Cox PH model for OS with growth and log-RD as covariates in Myeloma. 194

6.7 Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of being cured versus the growth rate from Weibull mixture model and PTC model in the Myeloma dataset . . . 198

6.8 Gelman plots showing the PSRF values across iterations in the mixture model for the cured proportion in the Myeloma dataset. Plotted on the first row are the factor loadings, coefficients of growth on π (alpha1) and λ (beta1) are on the second row, while the third row shows Gelman plots for the intercepts in the measurement models. 199

6.9 Gelman plots showing the PSRF values across iterations in the PTC model for the cured proportion in the Myeloma dataset. Plotted on the first row are the factor loadings, coefficients of growth on π and λ are on the second row, while the third row shows Gelman plots for the intercepts in the measurement models. 200

6.10 Histograms depicting the posterior distribution of the growth rate adjusted for log-RD as well as scatter plots of the probability of being cured versus the log-RD adjusted growth rate from Weibull mixture and PTC models in the Myeloma dataset 205

6.11 Gelman plots showing the PSRF values across iterations in the mixture model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the TTR. Plotted on the first row are the coefficients of growth (alpha1) and log-RD (alpha2) on π, on the second row are the factor loadings, while the third row shows Gelman plots for the coefficient of growth (beta1) and log-RD (beta2) in the TTR model. 206

6.12 Gelman plots showing the PSRF values across iterations in the PTC model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the TTR. Plotted on the first row are the coefficients of growth (alpha1) and log-RD (alpha2) on π, on the second row are the factor loadings, while the third row shows Gelman plots for the coefficient of growth (beta1) and log-RD (beta2) in the TTR model. 207

6.13 Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of attaining population OS versus the growth rate from Weibull population mixture and PTC models in the Myeloma dataset 211

6.14 Gelman plots showing the PSRF values across iterations in the population mixture model for the cured proportion in the Myeloma dataset with growth as a covariate on both π and the OS distribution. Plotted on the first row are the plots for the factor loadings, while coefficients of growth (alpha1) on π, and (beta1) on the OS are shown on the second row. Finally, the plots for the intercepts from the measurement models are on the third row. 212

6.15 Gelman plots showing the PSRF values across iterations in the population PTC model for the cured proportion in the Myeloma dataset with growth as a covariate on both π and the OS. Plotted on the first row are the plots for the factor loadings, while coefficients of growth (alpha1) on π, and (beta1) on the OS are shown on the second row. Finally, the plots for the intercepts from the measurement models are on the third row. . . . 213

6.16 Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of attaining population OS versus the growth rate adjusted for log-RD from Weibull population mixture and PTC models in the Myeloma dataset 216

6.17 Gelman plots showing the PSRF values across iterations in the population mixture model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the OS. Plotted on the first row are the intercept (alpha0), the coefficient of growth (alpha1) and log-RD (alpha2) on π, plots of the factor loadings tau[1], tau[2] and tau[2] are on the second row while the third row has plots for the intercept beta0, the coefficient for growth (beta1) and log-RD (beta2) in the OS model. . . 217

6.18 Gelman plots showing the PSRF values across iterations in the population PTC model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the OS. Plotted on the first row are the intercept (alpha0), the coefficient of growth (alpha1) and log-RD (alpha2) on π, plots of the factor loadings tau[1], tau[2] and tau[2] are on the second row while the third row has plots for the intercept beta0, the coefficient for growth (beta1) and log-RD (beta2) in the OS model. . . 218

6.19 Trace and density plots of the scale (lambda) and π (pi) from the Bayesian Weibull mixture model for TTR (A) and Bayesian Weibull population mixture model for OS (B) in the Myeloma dataset. 220

E.1 Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for TTR with 125 patients 257

E.2 Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for OS with 125 patients 258

F.1 Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for TTR in CLL fitted to data from 286 patients260

F.2 Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for OS with 301 patients 261

Symbols

Symbol	Description
$\boldsymbol{\alpha}$	Vector of covariates on the cured proportion
$\boldsymbol{\beta}$	Vector of covariates on the TTE
C	Random variable representing the censoring time
n	Number of individuals in a dataset
e_k	Number of events at the ordered time k in a Cox PH model
n_k	Number of individuals at tisk of the event just before the ordered time k
t	Observed time to event
T	Random variable representing time to event
$t_{(.)}$	Ordered time to event
δ	Censoring indicator
D	Observed data
$f(.)$	Probability density function
$F(.)$	Cumulative distribution function
$\forall$	For all
$h(.)$	Hazard function
$h_0(.)$	Baseline hazard function in Cox PH model
$h_p(.)$	Age-sex matched population hazard rate
$H(.)$	Cumulative hazard function
$H_0(.)$	Baseline cumulative hazard function in Cox PH model
i	Index for individuals in a sample
j	Index denoting observed variable, coefficient or time point
k	Number of observed random variables related to a latent construct in a measurement model
K	Number of years an individual was followed-up
I_j	Number of individuals at the start of year j
$I_j^{'}$	Effective number of individuals at risk in period j
d_j	Number of deaths observed in year j
w_j	Number of individuals who are still alive at the end of peroid j
p_j	Probability of dying in the j-th period
λ	Scale parameter in a TTE model or hazard ratio in Cox PH model
γ	Shape parameter in parametric survival model
$l(.)$	Log-likelihood function
$L(.)$	Log-likelihood function
p	Number of observed covariates per individual or p-value
$p(.)$	Density or pdf in Bayesian setting
π	Cured proportion
ρ	Mean number of remaining cancerous cells in a promotion time cure model
N	Number of cancer cells left after treatment
M_j	Random variable denoting time it takes the j-th cancer cell to grow to a detectable level
m	Number of imputations or imputed datasets
n_{cc}	Number of individuals in a complete case dataset
n_{ic}	Number of individuals with missing data
μ_j	Intercept in j-th measurement model
τ_j	Coefficient of ω in j-th measurement model
ϵ	Residual error in j-th measurement model
$Q(.)$	Expected value of log-likelihood in EM algorithm
r	Residual from TTE model
R	Set of individuals at risk of an event

Symbol	Description
s	Effective number of unknowns that can be estimated in a SEM
$S(.)$	Survivor function
$S_0(.)$	Baseline survivor function in Cox PH model
$S_p(.)$	Age-sex matched population survivor function
$S_u(.)$	Survivor function for those not cured following treatment and therefore still at a risk of relapsing
$S_d(.)$	Survivor function for those not cured following treatment and therefore still at a risk of dying from the disease
$N_i(.)$	Counting process for failures in Bayesian Cox PH model
P_{t-}	Available data before time t in Bayesian Cox PH model
$I_i(.)$	Intensity of the counting process in Bayesian Cox PH model
$Y_i(.)$	Censoring indicator in Bayesian Cox PH model
$\boldsymbol{\theta}$	Vector of unknown parameters in a model
$q(.)$	Proposal density in Metropolis-Hastings algorithm
φ	Acceptance probability in the Metropolis-Hastings algorithm
x	Observed covariate on the time to event
X	Random variable representing covariate on the time to event
z	Observed covariate on the cured proportion or missing random variables
Z	Random variable representing covariate on the cured proportion or missing random variables
ρ_0, ρ_1	Hyper-parameters in the inverse-Gamma prior for factor loadings
$g(.)$	Function of the posterior distribution of $\boldsymbol{\theta}$
J	Total number of iterations from MCMC chain
$J^{'}$	Number of remianing samples from MCMC chain after removal of burn-in
$D^{'}$	Data from posterior predictive distribution in Bayesian model
$p_{\boldsymbol{D}}$	Effective number of parameters in MCMC
$\bar{\boldsymbol{D}}$	Posterior mean of the deviance
$D(\bar{\boldsymbol{\theta}})$	Deviance evaluated at the posterior mean of $\boldsymbol{\theta}$
$\boldsymbol{\omega}$	$q \times 1$ vector of latent variables in structural equation model
q	Number of latent variables in structural equation model
$\boldsymbol{y}$	$k \times 1$ vector of continuous variables related to the latent variables $\boldsymbol{\omega}$
$\boldsymbol{y}^*$	$k \times 1$ vector of continuous variables related to the latent variables $\boldsymbol{\omega}$
$\boldsymbol{\Upsilon}$	$k \times q$ matrix of factor loadings in measurement model
$\boldsymbol{\Psi_\epsilon}$	Diagonal matrix of the errors in the measurement equations
$\boldsymbol{\Phi}$	Covariance matrix of the latent variables $\boldsymbol{\omega}$
$\boldsymbol{\Omega}$	Matrix of latent scores with n rows
$\boldsymbol{\eta}$	Outcome latent variables in structural equation model
$\boldsymbol{\nu}$	Intercept in structural model
$\boldsymbol{B}$	$r \times q$ non-singular matrix relating outcome latent variables to independent latent variables $\boldsymbol{\omega}$
$\boldsymbol{\zeta}$	A matrix of the errors in the structural model
$G(t)$	Time to event outcome in structural model
$\boldsymbol{S}$	Sample covariance matrix
$\boldsymbol{\Sigma}$	Model covariance matrix in structural equation

List of abbreviations

Abbreviation	Meaning
AIC	Akaike's Information Criterion
ADMIRE	ADdition of Mitoxantrone to Improve REsponse to FCR chemotherapy in patients with CLL
ARCTIC	Attenuated dose Rituximab with ChemoTherapy In CLL
BIC	Bayesian information criterion
BUGS	Bayesian inference Using Gibbs Sampling
CC	Complete case
CDF	Cumulative distribution function
CI	Confidence interval
CLL	Chronic lymphocytic leukemia
CrI	Credible interval
df	Degrees of freedom
DIC	Deviance Information Criterion
EM	Expectation-Maximization
GRD	Gelman-Rubin diagnostic
IQR	Inter-quartile range
JAGS	Just Another Gibbs Sampler
K-M	Kaplan-Meier
LL	Log-likelihood
MA	Missing at random
MC	Monte Carlo
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
ML	Maximum likelihood
MRC	Medical Research Council
MRD	Minimal residual disease
NP	Non-parametric
ONS	Office for National Statistics
OS	Overall survival
PL	Partial likelihood
pdf	Probability density function
PH	Proportional Hazards
PSRF	Potential scale reduction factor
RCT	Randomised controlled trial
RD	Residual disease
SD	Standard deviation
SEM	Structural equation model
SPMC	Semi-parametric mixture cure
TTE	Time to event
TTR	Time to relapse
UK	United Kingdom
VH	Variable heavy

Chapter 1

Background

Between 2014 and 2016 there were, on average, just over 360,000 newly diagnosed cancer cases in the United Kingdom (UK) each year according to Cancer Research UK [1]. In 2016, 163,000 of the reported 525,000 deaths in the UK were from cancer. Blood cancers are the fifth most common type of cancer in the UK and include various forms of leukaemia, lymphoma and myeloma [2]. Survival following treatment among those diagnosed with cancer can be as high as 50%, range 2-98%, over 10 years or more based on the 2010-11 Cancer Research UK data and depends on the type of cancer.

Cancer therapy aims to eradicate the primary disease. For solid tumours, the patient is cured of the disease if all the tumour is removed. However, most treatments will leave some residual microscopic disease which is resistant to the treatment. Further courses of treatment, called adjuvant therapy have to be administered to eliminate this residual disease (RD), and to prevent a relapse or metastasis, where secondary tumours develop at other sites in the body [3]. Whether all the tumour is removed usually cannot be easily determined in practice. For example, to detect the presence of microscopic disease in solid tumours, a polymerase chain reaction technique can be used to determine if there are metastases and microscopic tumour cells circulating in the blood [4, 5]. In blood cancers, the minimal residual disease (MRD), which expresses the RD

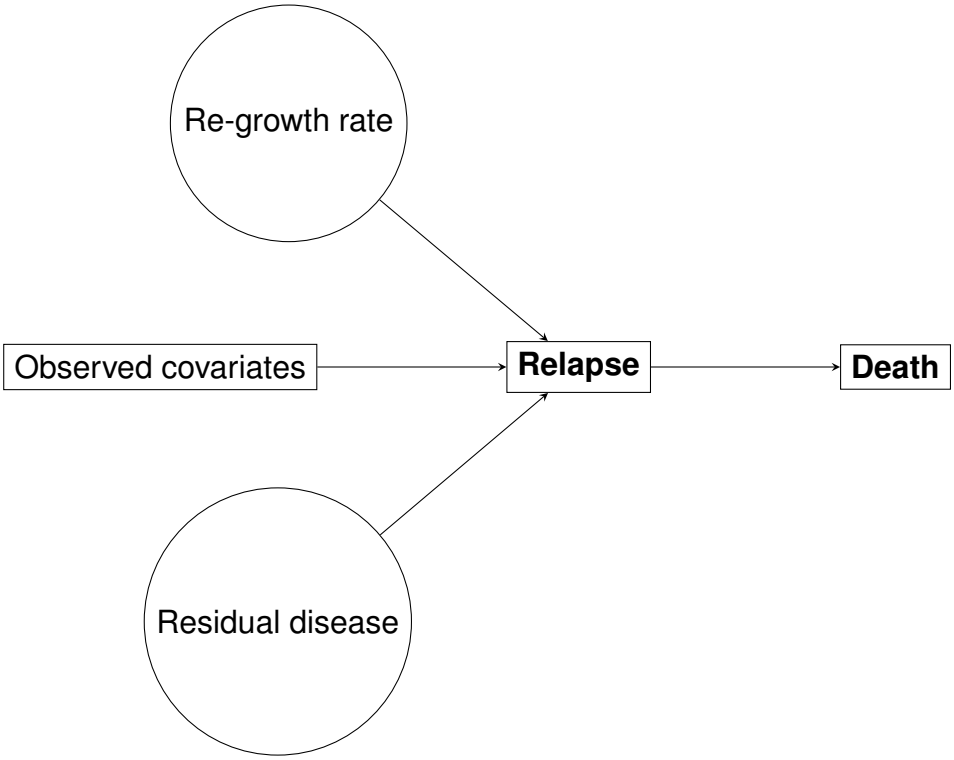


Figure 1.1: General model for time to relapse and subsequently death with both observed covariates and unobserved residual disease and tumour re-growth rate

as a percentage of tumour cells in the blood or bone marrow, where malignant cells can be seen in the blood but without clinical signs or symptoms, is one measure that can be used to indicate the amount of disease in the blood or bone marrow and hence to predict time to relapse (TTR) and overall survival (OS) in a group of patients [6, 7]. TTR and time to death will depend on both the amount of RD and the rate of re-growth of the primary tumour, as shown in Figure 1.1.

Advances in cancer treatment have resulted in some patients enjoying long-term disease-free status and better OS [8]. As a result, in some cancers, a proportion of patients will be considered 'cured' and will therefore not relapse following treatment, but will instead enjoy long-term OS that is comparable to that of disease-free individuals in the general population while the rest who are not cured, will relapse or die from

the malignancy [9]. The idea of a cure has been investigated in a number of solid cancers: colon, oesophageal and cervical among others [10, 11, 12, 13]. In Europe, specific proportions of those cured of lung, stomach, colon and rectal cancer have been reported [14]. In blood cancers, interest in investigating the possibility of a cure has been increasing [15]. Overall, an assessment of the trends of survival rates in different cancers has shown marked improvement in outcomes over time with estimated TTR curves showing plateaus indicating that a proportion of the patients might have been cured [16]. In most cases, the analysis of data from cancer patients focuses on TTR and/or OS and how they are related to important prognostic factors such as treatment, age, sex, stage (which describes the size of the tumour and its spread) and other biological markers of the disease where they are available [17]. However, the standard methods for analysis assume that all patients will either relapse or die from the disease given a sufficiently long follow-up period.

In order to accommodate the possibility of a cure following treatment, cure rate models are used to simultaneously estimate the proportion who are 'cured' following treatment and to model the TTR or OS for those who are not cured [18]. Cure in cancer can been defined in a number of ways: 1) the existence of a proportion of patients that will remain disease-free long-term after treatment often up to a pre-defined time post treatment, 2) cure in the clinical sense which involves the eradication of all malignant disease using various treatment methods, and 3) cure in the sense of having, within the patient population, a proportion whose OS rates are similar to that of individuals in the general population [19]. In most cases, cure has been defined as not relapsing or dying for 5 years after treatment. For slow-growing tumours, longer periods up to 10 years are required for cure to be considered [20].

Whether cure has been achieved cannot always be clinically defined and reported to the individual patient. However, the presence of a plateau in estimated TTR curves may lead to clinicians believing cure has been attained for a proportion of the patients.

With better treatments for cancer in general, has come interest in the investigation of cure in blood cancers including multiple myeloma [15, 21, 22, 23]. Assuming that cure is possible, cure rate models can be used to estimate 'cure' as defined in 1) and 3) above.

Where interest is in estimating the proportion of patients that will not relapse, plots of the TTR will level-off to a non-zero plateau, indicating the 'cured proportion'. This suggests that the population has two distinct groups; those that can be considered 'cured' or will not again relapse during their lifetime, and those who will eventually relapse or die from the disease or any other causes other than the disease. Patients who relapse can also get further causes of treatment including maintenance therapy through which they may end up enjoying long-term overall survival. Various formulations based on a logistic model for those that are cured, and parametric or semi-parametric models for those not cured have been extensively discussed in the literature [24, 25, 26, 27, 28]. Estimating the cured proportion by focusing on those who will not relapse following treatment can be challenging as such estimates are based on a particular sample of patients that might not represent the whole population with the disease. Since follow-up is limited, distinguishing between those who are cured and those who were just censored can be difficult.

The cure rate models can also be used to estimate the proportion whose OS or hazard of death returns to the same level as that expected in age-sex matched individuals in the general population [29, 30]. In these models, parametric distributions have often been used to model the OS for those not cured of the disease. The identifiability of the cure proportion itself and the parameters of the assumed distribution can be a challenge and will depend on the choice of the parametric distribution for the OS and the length of follow-up. Flexible parametric models that use splines to model the underlying hazard function can be more appropriate in these situations as they do not require specification of a parametric distribution for the OS among those not cured [31]. It is

also possible to specify a semi-parametric model for those who are not cured in these models.

The outcomes TTR, OS and the probability of a patient being cured of the disease usually depend on measurable prognostic factors such as age and treatment as well as measures of RD after treatment. However, relapse also depends on the rate of re-growth, which cannot always be measured and is not known at baseline. Where these variables are available, they have been shown to be useful for predicting outcomes. In breast cancer, RD measurements obtained through magnetic resonance imaging have been used to model the TTR [32]. In lung cancer, a similar approach has been used with serial measurements of tumour diameter used to work out the tumour growth rate [33]. For multiple myeloma and other blood cancers, a measure of the RD can be obtained by looking at the MRD. In the absence of the measures of RD and/or rate of re-growth, different methods have been used to estimate them, for example assuming latent exponential growth of the tumour [34]. The aim of these approaches was to aid in the design of clinical trials of different treatments based on the supposed understanding of the tumour (exponential) growth rate [35]. Other tumour growth rate distributions such as the lognormal and Gompertz have also been proposed [36, 37]. The growth of the RD following treatment [38], and how it relates to the effect of cytotoxic chemotherapy [39] has been shown to be important in explaining the chances of a relapse among cancer patients. Whilst these models are useful for designing trials, in the absence of data on RD and tumour re-growth rate, they rely on strong assumptions about the distributions of these measures which could not apply to all cancers.

Alternative methods that aim to infer the pattern of tumour growth and RD from the observed TTR, OS and the cured proportion have also been proposed [40, 41]. These methods, while attractive in that they directly obtain measures of the RD and growth from just the observed TTR or OS, are fraught with identifiability problems as several potentially correlated quantities are estimated from just the observed time a patient was

in the trial [42]. They also rely on strong assumptions about the models linking growth, RD and TTR or OS. Following on from this work, approaches for quantifying the effect of a treatment on the tumour were suggested [43, 44]. However, these were based on simulations of the tumour growth dynamics and how, based on this, the tumour would respond to treatment. The problem with this approach is that one cancer will likely be different to another in terms of tumour growth and other histopathological features making the application of such modelling approaches to treating the disease not easily generalizable [45]. Finally, another challenge is that while there is agreement on the benefit of modelling the tumour growth dynamics, such methods have not been widely incorporated in the design and analysis of clinical trials [46].

Rather than attempting to directly measure tumour growth following treatment, surrogate markers for tumour growth rate such as insulin-like growth factor indicators, can be used [47]. In pancreatic cancer for example, markers such as the epidermal growth factor, fibroblast growth factor and transforming growth factor beta receptors have been shown to be associated with aggressive tumour growth and therefore TTR and OS [48]. In some blood malignancies such as Myeloma and chronic lymphocytic leukaemia (CLL), measures of the RD can be obtained by measuring the MRD and using it as a surrogate for RD in prognostic models for the TTR and OS [7, 49, 50, 51].

The availability of potential surrogate measures of the unobserved (latent) tumour growth and RD make it possible to use structural equation models (SEMs) to model the role of tumour growth and RD on patient outcomes in cancer. SEM approaches are a part of a broader range of models that are used to validate relationships based on observed data [52], and can in this case, be used to model the relationship between the observed surrogate measures and the growth or RD and how these relate to the outcomes.

Originally, SEMs were used to model latent variables by assuming a multivariate normal distribution for the observed variables [53]. The methods have been extended

to situations where the observed variables are not normally distributed [54]. The relationship between latent variables and the observed data is modelled by examining the covariance structure of the proposed model and comparing it to that obtained from the observed data through the empirical covariance matrix [55]. As a result, most of the methods for fitting SEMs do not require individual observations of the latent variable. This feature makes SEMs attractive for investigating the role of measures that would not be directly observed. SEMs have been widely applied in the social sciences to model latent constructs, such as happiness, quality of life and intelligence. Recently, SEMs have been used to model the effect of a latent variable, difficulty in daily living, on health outcomes [56]. They have also been used to model the association of a latent variable, oxidative stress, with the risk of colorectal adenoma [57]. Building on this, the aim of this project is to further explore and develop methods for modelling the role of the often unobserved measures RD and tumour growth following treatment, on the TTR, OS and on the proportion cured.

When using these methods, decisions need to be made regarding the presence of potential observed variables that relate to the latent variable of interest and how the resulting latent variable can be measured [58]. Decisions on which observed variables to use would best be made by clinicians. However, this thesis will set out a general framework on how, given the presence of known surrogate measures of tumour growth or RD, SEMs can be used to model the effect of tumour growth and RD on TTR, OS and the probability of being cured. The framework will be demonstrated using data from UK academic clinical trials in Myeloma and CLL.

1.1 Aim

The aim of this project is to model the effect of tumour growth and RD on the TTR and OS and to estimate the role of these measures on the probability of being cured

following treatment, with application to Myeloma IX and CLL trial data.

Specific objectives

To investigate the role of available factors affecting TTR and OS following treatment using trial data, standard survival analysis methods that have been widely discussed already in the literature will be used. The proportions that will not relapse following treatment and whose OS is similar to individuals in the general population will be estimated using cure rate models. Finally, the effect of tumour growth and RD on the outcomes will be modelled using a combination of survival analysis and SEM techniques. Specifically, this project will:

1. Review standard models for time to event (TTE) data and their application to existing datasets, assessing goodness of fit and identifying limitations of these models.

2. Use maximum likelihood (ML) and Bayesian inferential approaches to estimate the proportion cured in 2 example datasets using a range of cure rate models and to compare the estimates.

3. Extend the current SEMs framework to include TTE and other outcomes to allow for the modelling of latent tumour characteristics, growth and RD, on outcomes in cancer.

4. Test the combined SEM-survival models on simulated data to assess how well these models fit the simulated data. This will include an investigation of the constraints and assumptions necessary for these models and how they can be fitted using maximum likelihood and Bayesian approaches.

5. Finally, apply the extended SEM models to the trial data to include the latent variables tumour re-growth and RD in modelling TTR, OS and the estimation of the proportion cured.

All methods discussed in this thesis will be applied to data from two trials in Myeloma and CLL conducted at the Clinical Trials Unit at the University of Leeds.

1.2 Outline of thesis

This chapter gives a brief description of the methods used to model outcomes in cancer trials in general and indicates how these can be extended to model the role of latent tumour measures on survival outcomes. The aim and objectives of this work are stated. The rest of the thesis is structured as follows. Chapter 2 gives a general review of the different methods used in analysing TTE data. A discussion on how parameters are estimated using both ML and Bayesian approaches is provided and is followed by graphical and other methods for assessing goodness of fit. Chapter 3 introduces the two example datasets used throughout this thesis and the application of standard TTE methods to the example data sets. In Chapter 4, cure rate models are used to estimate the cured proportion in these diseases. The SEM approach is introduced in Chapter 5, which also discusses extensions to TTE outcomes with applications to simulated data. The methods in Chapter 5 are again used to model the effect of tumour growth on the various TTE outcomes in the Myeloma dataset in Chapter 6. Finally, Chapter 7 reviews and compares strengths and limitations of all the approaches considered.

Chapter 2

Methods for analysing time-to-event data

2.1 Introduction

In this chapter, various methods for estimating parameters in TTE models are briefly discussed. In most analyses involving TTE data from randomised control trials (RCTs) in cancer, researchers are interested in the TTR or OS of a group of patients following treatment and how important prognostic factors influence these outcomes including randomised allocation. To achieve this, survival analysis methods are used to model the TTE, T. The role of important prognostic factors on T can be investigated using non-parametric and semi-parametric approaches, where no assumptions about the distribution of T are made [59, 60]. Alternatively, parametric survival models can be used where the role of the treatment and other factors on the TTE are modelled through one or more parameters governing the distribution of T [61]. These methods are briefly discussed in Sections 2.3, 2.4 and 2.5 respectively.

In instances where not all patients will relapse or die from the cancer following treatment, the usual TTE methods might not be appropriate as they assume that all patients will experience the event. In these situations, cure rate models have been proposed to simultaneously estimate the proportion of patients who will not relapse or die from

the cancer following treatment, and also to model the time, T, to relapse or death for those not cured by the treatment [62]. The cure rate models can also be used to estimate 'cure' as the proportion whose OS is similar to age-sex matched individuals in the general population who are without the disease. For this latter case, the 'population' cure rate models fitted in this thesis incorporate population survival data from the UK national mortality registers for 2015 - 2017 [63]. Model specifications for the cure rate models are discussed in Section 2.6 while the derivation of age-sex matched hazard and survival rates from the UK Office for National Statistics (ONS) data is shown in Section 2.7.1. We start by presenting a general survival analysis.

2.2 Overview of TTE data

A feature with TTE data is that over the observation period, not all individuals will experience the event. These individuals are described as 'censored' if they drop out of the study, or if the study ends before they experience the event. In the analysis, they are counted event-free during the study. Throughout this thesis, censoring will be assumed independent of the event of interest. In our set-up we assume that U_i is the TTE for the i-th subject and C_i is the censoring time for the i-th individual. We define a censoring indicator δ_i depending on whether the i-th individual had the event as

$$\delta_i = \begin{cases} 1 & \text{if } U_i < C_i \\ 0 & \text{Otherwise} \end{cases}$$

Treating failure and censoring as competing risks, we obtain $T_i = min(U_i, C_i)$ as the TTE for the i-th individual. The analysis then focuses on the pairs (t_i, δ_i) where t_i is the observed TTE.

The analysis of TTE data usually focuses on two summaries, the survival and hazard functions from which other functions may be obtained [64]. Let $T \geq 0$ be a random

variable representing the TTE. If T is continuous, then it has a probability density function (pdf), $f(t)$, and from this we can evaluate the cumulative distribution function (CDF) of T which is defined as the cumulative probability of an event occurring before time t i.e.

$$F(t) = \mathsf{Pr}(T < t) = \int_0^t f(u)du \tag{2.1}$$

The survivor function $S(t)$ is defined as the probability that the TTE is at least t, that is

$$S(t) = \mathsf{Pr}(T \geq t) = 1 - F(t). \tag{2.2}$$

The hazard function is defined as the instantaneous probability of an event occurring at a time t, conditional on survival up to the time t. It can be expressed in terms of the pdf, $f(t)$, and the survivor function $S(t)$ as

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.3}$$

The cumulative hazard of an event occurring by time t can be obtained from the hazard and survivor functions as

$$H(t) = \int_0^t h(u)du = -\log S(t). \tag{2.4}$$

The hazard, cumulative hazard and survivor functions can then be used to model the observed data. Inference is made based on the likelihood function which is the joint probability distribution of a sample of observed data as a function of some unknown parameters of interest. With TTE data, the likelihood function based on data from the n individuals is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\boldsymbol{\theta})^{\delta_i} S(t_i|\boldsymbol{\theta})^{1-\delta_i}, \tag{2.5}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters that govern the distribution of the TTE. Alternatively, the likelihood can be expressed in terms of the hazard and survivor functions as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} h(t_i|\boldsymbol{\theta})^{\delta_i} S(t_i|\boldsymbol{\theta}). \tag{2.6}$$

Covariates can be used to model the shape of the hazard in semi-parametric and parametric survival models as will be discussed in detail later.

2.3 Non-parametric approaches

The simplest method for analysing TTE data is the non-parametric (NP) approach which does not impose any assumptions on the distribution of the TTE. The most commonly used approach is the Kaplan-Meier (K-M) estimate of the survival function which is based on the empirical data as follows. Let $0 < t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ be the ordered observed times to the event not including the censored times. Further, let e_k be the number of events at time $t_{(k)}$ and let n_k be the number of individuals at risk of the event just before time $t_{(k)}$. Based on this, the K-M estimate of the survival function is the product

$$\hat{S}(t) = \prod_{k:t_{(k)}<t} \left(1 - \frac{e_k}{n_k}\right) \tag{2.7}$$

where in (2.7), $\frac{e_k}{n_k}$ represents the probability of having the event at time $t_{(k)}$ given the individual is at risk just before $t_{(k)}$. The K-M estimate is a step function characterised by jumps at the observed event times. If there is no censoring, the K-M estimate coincides with the empirical survivor function which is itself an estimate of the survivor function that does not depend on distributional assumptions about T. If the largest observation is censored, the estimate is undefined beyond the last event time. Comparison of survival

functions among groups of individuals can be made using the log-rank test [65]. This method is useful for graphically comparing survival among groups.

2.4 Semi-parametric approaches

Although the NP methods work well in exploratory analysis, they cannot be used in the presence of several explanatory variables; instead, semi-parametric approaches can be used to summarise TTE data by estimating the hazard and survivor functions without making strong assumptions about the baseline hazard function. The most widely used semi-parametric survival model is the Cox proportional hazards (PH) model [66].

2.4.1 The Cox PH model

The Cox PH model has become the most widely used method for analysing data with TTE outcomes. In this model, no assumptions are made regarding the distribution of the baseline hazard function when investigating the role of covariates on the TTE. The Cox PH model assumes a multiplicative effect of each covariate on the unspecified hazard function that is constant over time. As an example, suppose that t_i is the TTE for an individual i. Then the hazard function for this individual is

$$h(t_i) = h_0(t_i)\lambda_i, \ \ \forall\, t_i > 0 \tag{2.8}$$

where $h_0(t_i)$ is the unspecified baseline hazard and λ_i is the multiplicative factor for individual i. If for an individual, i, we observe a vector of p covariates $\boldsymbol{x_i}$, then λ_i can be expressed in terms of covariates as

$$\lambda_i = \exp(\boldsymbol{\beta^T x_i}) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}).$$

2.4.2 Partial likelihood for the Cox PH model

To obtain the parameter estimates in the Cox PH model, a partial likelihood approach proposed by Cox [67] is used. At each failure time, let $R_j = \{i, i = 1, \ldots, n : T_i \geq t_{(j)}\}$ be the set of all patients at risk of the event together with their observed covariate values $\boldsymbol{x_i}$. Assuming a single patient has the event at time $t_{(j)}$, given R_j were at risk just prior to time $t_{(j)}$, the probability that subject i had the event is

$$\frac{h_0(t_{(j)})\lambda_i}{\sum_{k\in R_j} h_0(t_{(k)})\lambda_k} = \frac{h_0(t_{(j)})\exp(\boldsymbol{\beta^T x_i})}{\sum_{k\in R_j} h_0(t_{(k)})\exp(\boldsymbol{\beta^T x_k})}$$

and since $h_0(t)$ is common and can be cancelled out, this is just

$$\frac{\exp(\boldsymbol{\beta^T x_i})}{\sum_{k\in R_j}\exp(\boldsymbol{\beta^T x_k})}.$$

The partial likelihood (PL) is thus based only on the observed covariate values

$$PL(\boldsymbol{\beta}) = \prod_j \frac{\exp(\boldsymbol{\beta^T x_{i(j)}})}{\sum_{k\in R_j}\exp(\boldsymbol{\beta^T x_{k(j)}})} \tag{2.9}$$

where $i(j)$ refers to the individual who had the event at time $t_{(j)}$. Its logarithm is

$$pl(\boldsymbol{\beta}) = \sum_j \left(\boldsymbol{\beta^T x_{i(j)}}) - \log\left\{ \sum_{k\in R_j} \exp(\boldsymbol{\beta^T x_{k(j)}}) \right\} \right).$$

Differentiating this with respect to $\boldsymbol{\beta}$

$$\frac{\partial pl(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_j \left(\boldsymbol{x_{i(j)}}) - \frac{\sum_{k\in R_j} \boldsymbol{x_{k(j)}}\exp(\boldsymbol{\beta^T x_{k(j)}})}{\sum_{k\in R_j}\exp(\boldsymbol{\beta^T x_{k(j)}})} \right) \tag{2.10}$$

and equating to $\mathbf{0}$ gives the estimates of $\boldsymbol{\beta}$.

In some cases, more than one individual has the event at time $t_{(j)}$. The PL (2.9) can be modified to take this into account by introducing a failure set, F_j in the numerator to

the PL to represent all individuals who had the event at time $t_{(j)}$.

2.5 Parametric models

While semi-parametric approaches are attractive in that the baseline hazard function does not have to be specified when modelling the effect of covariates on the TTE, it may sometimes be necessary to estimate the actual baseline hazard. In parametric models, the hazard function is estimated by assuming a distribution to represent the underlying population. If the assumed distribution is correct, the parametric models have an advantage over the semi-parametric approach as it is possible to estimate the distribution of survival time using the full likelihood unlike having to use the partial likelihood as in semi-parametric methods. Further, if the correct form of the parametric distribution is specified, parametric models are more efficient and will give more precise estimates of the covariate effects. Residuals from parametric models take the familiar form as the difference between observed and estimated values. The main disadvantage with parametric models is that they rely on the assumption that the correct underlying population distribution has been specified. This makes semi-parametric models such as the Cox PH model more attractive in practice as it is possible to model the effect of covariates on the TTE without having to worry about the underlying distribution of the hazard function in the population.

Parametric models follow a specified probability distribution based on a number of fixed parameters [68]. Parameters may represent the hazard function and how it varies with time, for example. Explanatory variables of the TTE can be modelled through one or more parameters of the chosen parametric distribution. Common parametric survival distributions are Exponential, Weibull and Gompertz among several others. In this thesis we focus on the Weibull and Exponential distributions when referring to parametric models. The likelihood for the parametric models is similar to (2.5) where, for

example the parameter vector $\boldsymbol{\theta}$ includes the shape and scale parameters if the Weibull model is specified, both of which may be functions of the observed covariates, or it may represent the rate parameter if the Exponential model is used. As an example, the scale parameter of a Weibull model may be related to the observed covariates through

$$\lambda_i = \exp(\boldsymbol{\beta^T x_i}) = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}). \tag{2.11}$$

for $i = 1, 2, \ldots, n$ if we observe data from n individuals. Here, β_0 is the intercept and $\beta_1, \ldots, \beta_p$ are the coefficients relating each of the p observed covariates to the TTE. We next briefly describe Exponential and Weibull distributions used to fit parametric TTE models.

Exponential and Weibull distributions

The Exponential distribution is the simplest parametric TTE model and it assumes a constant hazard of the event over time. The Exponential model can be specified in terms of the pdf for individual i as follows

$$f(t_i) = \lambda_i \exp(-\lambda_i t_i) \tag{2.12}$$

where λ_i is as defined in (2.11). From this, the hazard and survivor functions can be obtained as $h(t_i) = \lambda_i$ and $S(t_i) = \exp(-\lambda_i t_i)$, respectively.

The Weibull distribution is one of the most widely used distributions in analysing TTE data. In its most general form, it is equivalent to other distributions, for specific values of the shape parameter, γ. We consider the 2-parameter Weibull distribution model. The pdf can be parameterised as

$$f(t) = \lambda \gamma t^{\gamma - 1} e^{-\lambda t^{\gamma}} \tag{2.13}$$

where $\lambda > 0$ is the scale parameter and $\gamma > 0$ describes the shape of the distribution. The hazard function can be expressed as

$$h(t) = \lambda\gamma t^{\gamma-1} \tag{2.14}$$

which is monotonically increasing if $\gamma > 1$, constant if $\gamma = 1$ (in which it is equivalent to the Exponential distribution) and decreasing if $\gamma < 1$. The corresponding survivor function is

$$S(t) = \exp(-\lambda t^{\gamma}).$$

The scale parameter can be parametrised in terms of covariates and regression parameters whereas the shape γ, is mostly assumed fixed. To model the effect of covariates on the scale parameter for the i-th individual, we can use (2.11) to obtain the hazard for individual i at time t_i

$$h(t_i) = \lambda_i\gamma t_i^{\gamma-1}. \tag{2.15}$$

Inference can then be done by estimating the parameters given data from n individuals using by maximising the likelihood functions (2.5) and (2.6) using methods that will be discussed in Section 2.9.

2.6 Cure rate models for the proportion that will never relapse following treatment

To estimate the proportion that will not relapse, further modifications of the likelihood can be made. In this case, the survivor function is given as

$$S(t) = \pi + (1-\pi)S_u(t) \tag{2.16}$$

where π represents the proportion that is cured following the treatment and will remain relapse-free, $f_u(t)$ the density and $S_u(t)$ represents the survivor function of those who are not cured following the treatment and will therefore experience a relapse. This type of cure rate model is referred to as a mixture model. From (2.16), the pdf is given by

$$f(t) = (1-\pi)f_u(t)$$

while the hazard is

$$h(t) = \frac{(1-\pi)f_u(t)}{\pi + (1-\pi)S_u(t)}$$

The resulting likelihood has the same form as (2.5) and (2.6).

2.7 Cure rate models for estimating the proportion with OS similar to the general population

To estimate the proportion with OS similar to age-sex matched individuals in the general population, a further modification of (2.16) can be made to incorporate the population survival rates to get

$$S(t) = S_p(t) * (\pi + (1-\pi)S_d(t)). \tag{2.17}$$

Here $S_p(t)$ is the survival probability of a sex-aged matched individual in the general population at the time at which the event is observed in the individual with the disease and $S_d(t)$ is the survivor function for those who will eventually die from the disease. The hazard and pdf can then be derived as before to build the likelihood function. The populations survival or hazard rates are briefly discussed next.

2.7.1 Calculating the population hazard and survival rate

To fit the models in Section 2.7, the population hazard rate $h_p(t)$ for each individual in the dataset has to be estimated, matched for age and sex at the end of that individual's follow-up using the life-table approach in the following manner. Suppose there are I_j individuals in the study at the start of period j which could be 1 month, 1 year, etcetera. Suppose d_j deaths are observed in period j, then we can calculate the number of censored individuals w_j in this period using

$$w_j = I_j - d_j.$$

Then the effective number of people at risk in the j-th period is

$$I_j^{'} = I_j - \frac{w_j}{2}$$

since under constant within-interval hazard, at least 50% of those censored will be at risk for at least half the time [69]. Then, the probability of dying in the j-th period is

$$p_j = 1 - \frac{d_j}{I_j^{'}}$$

and we can denote this by $S_p(t)$, the survival probability at time t. If the individual in the study is followed-up for $j = 1, 2, \ldots, K$ periods, then

$$S_p(t) = \prod_{j=1}^{K} p_j. \tag{2.18}$$

The hazard rate can then be obtained from (2.18) as

$$h_p(t) = -\log(S_p(t)). \tag{2.19}$$

2.8 Bayesian TTE models

In some situations, a Bayesian approach can be useful, since it allows additional external information about the parameters of interest to be incorporated as prior information. This information could be in the form of data from previous studies, or expert knowledge from those familiar with the field of study. In a Bayesian setting, inference is made based on the joint posterior distribution of the parameters of interest given the observed data. The observed data is distributed according to the likelihood $p(D|\boldsymbol{\theta})$ where D could be a vector or matrix of observed TTEs, censoring indicators or covariates while $\boldsymbol{\theta}$ is a vector of unknown parameters which are assumed to be random. The researcher also specifies a prior distribution of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$. The posterior is then calculated as

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.20}$$

Note that the likelihood $p(D|\boldsymbol{\theta})$ can be from any of the specified semi-parametric, parametric and structural equation models. The fitting of Bayesian models in general is discussed in Section 2.9.3.

2.9 Overview of parameter estimation in TTE models

Common approaches that exist for fitting models to TTE data are based on maximum likelihood (ML). These methods are straight-forward to implement when fitting parametric survival models. For semi-parametric models, it may not be possible to directly estimate parameters from the full likelihood using ML techniques. Instead, the partial likelihood first proposed by Cox [67], is used to estimate parameters in the semi-parametric models. To fit cure rate models using ML techniques, the Expectation-Maximisation (EM)algorithm can be used [70]. The EM algorithm is an iterative procedure for max-

imising likelihood functions that depend on unobserved or latent variables. As the cure rate models aim at estimating the proportion cured which is itself not observed, the EM algorithm has been used to estimate parameters in these models. Finally, by specifying prior distributions on the parameters of the chosen model, we can estimate their posterior distributions given the observed data to fit both standard and cure rate models using Markov Chain Monte Carlo (MCMC) methods in a Bayesian setting. For the semi-parametric TTE models, a counting process is adopted in the Bayesian MCMC implementation. Each of these approaches are next discussed in turn below.

2.9.1 ML for fitting standard models

ML estimates are found by maximising the likelihood function which is the probability of observing the data D given some unknown, but fixed parameters $\boldsymbol{\theta}$. Standard statistical packages such as Stata and R can be used to obtain parameter estimates in these models [71, 72]. ML estimates of the parameters given a parametric distribution for T can be obtained by maximising the log-likelihood (LL) of the data D given the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, $l(D|\boldsymbol{\theta})$. The ML estimates of of the elements in $\boldsymbol{\theta}$, $\hat{\theta}_j, j = 1, 2, \ldots, p$ can be obtained from the score functions

$$\frac{\partial l(D|\hat{\theta}_j)}{\partial \hat{\theta}_j} = 0 \tag{2.21}$$

To measure the amount of information about the unknown parameters in the data, we work out the Fisher information matrix which gives the variances of the scores. Differentiating again and taking expectations provides an estimate of the variance-covariance matrix

$$Var(\boldsymbol{\theta}|D) = \begin{bmatrix} -E\left\{\frac{\partial^2 l(D|\boldsymbol{\theta})}{\partial \theta_1^2}\right\} & \cdots & -E\left\{\frac{\partial^2 l(D|\boldsymbol{\theta})}{\partial \theta_1 \theta_p}\right\} \\ \vdots & \ddots & \vdots \\ -E\left\{\frac{\partial^2 l(D|\boldsymbol{\theta})}{\partial \theta_p \theta_1}\right\} & \cdots & -E\left\{\frac{\partial^2 l(D|\boldsymbol{\theta})}{\partial \theta_p^2}\right\} \end{bmatrix} \tag{2.22}$$

where the diagonal elements represent the variances of $\theta_1, \theta_2, \ldots,$ and θ_p respectively [73]. Inference can then be made based on these estimates.

2.9.2 Using the EM algorithm

The EM algorithm provides an alternative way to find parameter estimates using ML in situations where the model includes latent variables or to deal with missing data. The EM algorithm can thus be used to find estimates in cure rate models by assuming that for each individual, there is an unobserved latent variable representing the probability that they are cured or not. The EM algorithm is briefly described next.

Suppose on the complete data $\boldsymbol{D}$, we specify the model $p(\boldsymbol{D}|\boldsymbol{\theta})$. The complete data $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{Z})$ includes the observed data $\boldsymbol{X}$ and the latent or missing variables $\boldsymbol{Z}$. The complete data likelihood is $L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z})$. The unknown parameters $\boldsymbol{\theta}$ can be estimated by maximising the marginal likelihood of the observed data

$$L(\boldsymbol{\theta}|\boldsymbol{X}) = \int L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}. \tag{2.23}$$

In most cases, the integral in (2.23) is not tractable. The EM algorithm is thus used to maximise the marginal likelihood by iteratively applying a two-step process. The first step is the expectation step (E step). At the E step, work out

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) = E_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(j)}} \log L(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Z}) \tag{2.24}$$

which is the expected value of the LL function of $\boldsymbol{\theta}$, with respect to the current conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ and the current estimates of the parameters $\boldsymbol{\theta}^{(j)}$. This is followed by the maximisation step (M step) which involves working out parameters that maximise $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{p})$,

$$\boldsymbol{\theta}^{(j+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}) \tag{2.25}$$

For this work, the latent variable Z will represent an indicator for whether a patient is cured of disease or not. Details of how the EM algorithm can be use in general [74], and how they can be used in estimating parameters in cure rate models have been discussed elsewhere [25, 27]. The model fitting follows the following steps:

1. Set the unknown parameters θ to some initial random values
2. Work out the probability of each possible value of the unknown latent variable Z given θ using (2.24)
3. Work out an improved estimate of θ using the information about Z obtained in step 2 using (2.25)
4. Repeat steps 2 and 3 until convergence

There are several advantages of the EM algorithm, the most important in our context is the ability to fit models even if we have missing data or latent variables, which is sometimes not possible with ML techniques. Moreover, each iteration is guaranteed to improve the likelihood until the maximum is reached, and convergence will be fast if analytical expressions for the parameters at the M-step can be found. Finally, the maximisation does not involve working out derivatives leading to easier programming. The main disadvantage is that convergence is generally slow if no analytical expressions for the parameters are available at the M-step, making it necessary to use numerical methods that further slow down the maximisation process [75]. Convergence is also in doubt when there is a large percentage of missing data or we have high dimensional data. The EM algorithm will be used to estimate parameters from mixture models in Chapter 4.

2.9.3 Using Markov Chain Monte Carlo methods to fit Bayesian models

Based on the posterior (2.20), MCMC methods can be used to estimate the parameters in the model. The MCMC algorithm is attractive because, with only mild conditions, it makes it possible to draw samples from a joint posterior distribution. This is made possible by iteratively drawing samples from conditional distributions of parameters until convergence. It is thus feasible to draw samples from complicated joint distributions with high dimensionality, something which might not be possible with ML techniques. MCMC methods are relatively easy to implement as long as we can specify conjugate prior distributions of the parameters in the model, which are combined with the observed data via the likelihood, to work out the joint posterior distribution. One challenge with MCMC is that it is difficult to assess accuracy and convergence. As a result of this, there is a danger of making incorrect inferences based on an MCMC algorithm that may not have converged [76]. Another disadvantage of MCMC methods is that of computational cost when the analysis involves evaluating intractable integrals in complex models which may result in the algorithm taking a long time to converge [77]. There are several tools used for checking convergence of MCMC algorithms that will be discussed in Section 2.10.4.

The Gibbs sampler

One of the commonly used MCMC techniques is the Gibbs sampler. Suppose D represents the observed data and we are interested in estimating the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. A sufficiently large sample from the joint posterior distribution $p(\boldsymbol{\theta}|D)$ can be obtained from the Gibbs sampler algorithm as follows. At the $(j+1)$-th iteration with current values $\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_p^{(j)}$:

1. Generate $\theta_k^{(j+1)}$ from $p(\theta_k|\theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j)}, \ldots, \theta_p^{(j)}, D)$
2. Generate $\theta_{k+1}^{(j+1)}$ from $p(\theta_{k+1}|\theta_1^{(j)}, \ldots, \theta_k^{(j+1)}, \theta_{k+2}^{(j)}, \ldots, \theta_p^{(j)}, D)$

3. ⋮

4. Generate $\theta_p^{(j+1)}$ from $p(\theta_p|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{p-1}^{(j+1)}, D)$

This process is repeated many times until the joint posterior converges to a stationary distribution. Inferences can then be made using the parameter values of $\boldsymbol{\theta}$ in this stationary distribution.

There are some problems with using the Gibbs sampler. For example, it works well when it is easy to sample from the full conditional distributions although but this might not always be the case. Another challenge is that the algorithm may take long to converge due to slow mixing if some of the parameters are highly correlated.

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm can be used to sample from the joint posterior in situations where the posterior distribution does not take the form of a known conjugate distribution or where the full conditionals do not resemble known distributions. When this is the case, the Gibbs sampling algorithm might not work and so it is more appropriate to use the Metropolis-Hastings algorithm. The algorithm is set-up as follows. Given the full posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$,

1. Choose some arbitrary starting values for the unknown parameters $\boldsymbol{\theta}^{(0)}$

2. At the j-th iteration, draw $\boldsymbol{\theta}^*$ from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})$

3. Calculate the acceptance probability

$$\varphi = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})/q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(j-1)})}{p(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{x})/q(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^*)}$$

4. Accept the current value of $\boldsymbol{\theta}^j$ as $\boldsymbol{\theta}^*$ with probability $min(\varphi, 1)$, otherwise, set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$

Once again, repeat this process until $p(\boldsymbol{\theta}|\boldsymbol{x})$ converges to a stationary distribution.

In both Gibbs and Metropolis-Hastings algorithms, the MCMC chain converges to a stationary distribution under mild conditions which is itself a random sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}$ where J is the total number of iterations for which the MCMC algorithm was run [78]. The algorithm is normally run by supplying initial values of the chain $\boldsymbol{\theta}^{(0)}$. Since the chosen initial values may greatly differ from the resulting stationary distribution, they are removed from the chain such that they do not influence the resulting posterior summaries. This is called the burn-in period. Researchers would normally run multiple chains with different starting values to ensure convergence to the correct distribution. As the final MCMC sample itself is not independent it is necessary to monitor the autocorrelations of the generated values and retain only certain values with a given lag. This is called thinning and involves discarding all but the r-th sampled values for example where $r > 1$ [79].

Bayesian Cox PH models

To fit a Cox PH model in a Bayesian setting, a counting process approach is used in order to deal with the unspecified baseline hazard $h_0(t)$ [80]. This makes the estimation of the baseline hazard and regression parameters through MCMC methods possible in the following manner. Suppose we observe data from $i = 1, 2, \ldots, n$ individuals, and that $N_i(t)$ is a process that counts all failures that have occurred up to time t and $I_i(t)$ is an intensity process given by

$$I_i(t)dt = E(dN_i(t)|P_{t-}) \tag{2.26}$$

where $dN_i(t)$ is the increment in N_i over the small interval $[t, t+dt)$, and P_{t-} represents the available data before time t [81]. The expectation in (2.26) can be looked at as the probability that an individual i will have the event in the interval and when $dt \to 0$. This

is the instantaneous hazard for the i-th subject at time t and can thus take the PH form

$$I_i(t) = Y_i(t)h_0(t)\lambda_i \tag{2.27}$$

where (2.27) is now the familiar Cox PH model format with $\lambda_i = \exp(\boldsymbol{\beta^T x_i})$ and $Y_i(t)$ is an observed process takes the value 1 if individual i is under observation or at risk at time t and 0 if not.

The observed data is the set $D = \{N_i(t), Y_i(t), \boldsymbol{x_i}; i = 1, 2, \ldots, n\}$ and we are interested in the unknown parameters $\boldsymbol{\beta}$ and the cumulative hazard $H_0(t) = \int_0^t h_0(u)du$. Given the observed data D and by specifying priors for our unknown parameters, we can express the posterior distribution as

$$p(\boldsymbol{\beta}, H_0(t)|D) \propto p(D|\boldsymbol{\beta}, H_0(t))p(\boldsymbol{\beta}|H_0(t))p(H_0(t)). \tag{2.28}$$

Under non-informative censoring, the likelihood of the data is proportional to

$$\prod_{i=1}^{n}\left[\prod_{t\geq 0} I_i(t)^{dN_i(t)}\right]\exp\left(-\int_{t\geq 0} I_i(v)dv\right). \tag{2.29}$$

It now looks like the increments $dN_i(t)$ in the small interval $[t, t + dt)$ are independent Poisson random variables with means $I_i(t)dt$ i.e.

$$dN_i(t) \sim \text{Poisson}(I_i(t))dt$$

.

We may then write

$$I_i(t) = Y_i(t)\lambda_i dH_0(t)$$

where $dH_0(t)dt = H_0(t)dt$ is the increment or jump in the integrated baseline hazard function during the time interval $[t, t + dt)$. As the conjugate prior for the Poisson mean

is the Gamma distribution, we can assume

$$dH_0(t) \sim \text{Gamma}(cdH_0^{\star}(t), c) \tag{2.30}$$

where $dH_0^{\star}(t)$ is the prior guess of the unknown hazard function and c is some measure of our confidence in this guess. Small values of c would imply weak prior beliefs. What remains is to then choose an appropriate prior for $\boldsymbol{\beta}$ to work out the posterior distribution.

Posterior summaries and inference based on the MCMC output

The final MCMC output is now a random sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J')}$ of the joint posterior distribution of the $\boldsymbol{\theta}$'s, where J' is the number of remaining samples with the burn-in removed. Using this sample, the measures of interest can be calculated as functions of the parameters of interest. More specifically, the posterior mean and standard deviation (SE) for a function of $\boldsymbol{\theta}$, $g(\boldsymbol{\theta})$ can be obtained from

$$\widehat{E}(g(\boldsymbol{\theta}|D)) = \overline{g}(\boldsymbol{\theta}) = \frac{1}{J'} \sum_{j=1}^{J'} g(\boldsymbol{\theta}^{(j)}) \tag{2.31}$$

and

$$\widehat{SD}(g(\boldsymbol{\theta}|D)) = \sqrt{\frac{1}{J'-1} \sum_{j=1}^{J'} \left\{ g(\boldsymbol{\theta}^{(j)}) - \widehat{E}(g(\boldsymbol{\theta}|D)) \right\}^2} \tag{2.32}$$

respectively. In a similar manner, other measures of interest such as the posterior median and quantiles from which 95% credible intervals (CrIs) may be obtained, can be calculated.

2.10 Assessing the model fit in TTE models

The fitted model can be assessed to see how well it fits the data or, where the Cox PH model is used, whether it violates the proportional hazards assumption, before any inference is made. For TTE data, it is not possible to use standard residuals to assess model fit due to censoring. As a result, various methods for assessing proportional hazards models have been proposed and we discuss them next.

2.10.1 Schoenfeld residuals

One way to assess the PH assumption is to use Schoenfeld residuals [82]. Schoenfeld residuals are calculated as the difference between the observed covariate value for the i-th subject and its conditional expectation given at any event time $t_{(j)}$. It is given by

$$r_j(\boldsymbol{\beta}) = \boldsymbol{x_i} - \sum_{m \in R_j} \frac{\boldsymbol{x_m} \exp(\beta^T x_m)}{\sum_{k \in R_j} \exp(\beta^T x_k)} \tag{2.33}$$

where β is a coefficient from the PH model. The form given in (2.33) is the contribution of a given event to the score function of the partial likelihood such that $\sum_j \hat{r}_j(\hat{\boldsymbol{\beta}}) = 0$. The PH assumption is then checked by plotting the Schoenfeld residuals against time for each covariate. If the PH assumption is valid, the residuals have a mean 0 over the whole range of the observed times. If, however, the plot shows deviations from 0 over time, it might imply that the hazards are not proportional and that there is a time dependent effect that has to be taken into account [83]. In practice, scaled Schoenfeld residuals which are obtained by transforming the residual to have an approximate variance of 1 are used. This makes it possible to formally test hypotheses such as the PH assumption and procedures for doing there are available in standard statistical software.

2.10.2 Other residuals and methods for assessing model fit

Martingale residuals can also be useful when interest is on investigating whether the correct functional form of a continuous covariate has been given. A plot of these residuals against the covariate will show a smoothing line close to 0 if the functional form is correct [84]. Alternatively, Cox-Snell residuals can be used to perform goodness-of-fit tests for proportional hazards survival models (as well as other survival models) by plotting the Cox-Snell residuals against the cumulative hazard function [85]. Good fit will be depicted by a straight line through the origin with gradient 1. One problem with the Cox-Snell residuals is that the plot might depict a linear relationship even for models that do not satisfy the PH assumption. Because of the limitations of the Cox-Snell residuals and the situations which my require use of Martingale residuals not encountered in this work, Schoenfeld residuals were used to assess the PH assumption in both univariable and multivariable models in this work.

For Exponential and Weibull TTE models, goodness of fit may be assessed informally by looking at how the survivor function $S(t)$ or some function of it varies with the TTE or some function of t. To assess whether the PH assumption holds for example, a plot of

$$\log(-\log(\hat{S}(t_i))) \text{ against } \log(t_i)$$

where $\hat{S}(t)$ is the K-M estimate can be used [86]. If the PH holds, lines for different levels of a categorical variables will be parallel for the survival model including parametric models such as the Exponential and Weibull models. Another way for assessing goodness of fit in parametric models is to plot the fitted survival curve based on the model parameters alongside the K-M plot to see how closely the fitted model matches the empirical data.

2.10.3 Assessing model fit in cure rate models

There are no universally used methods for assessing fit in cure rate models such as Schoenfeld residuals. Pseudo-residuals have been proposed for assessing model fit in the distribution of T for those not cured [87]. Other methods have focused on testing the presence of the cured proportion itself (π) and ignored the role of covariates [88], while other proposed methods test whether the cured proportion as a function of covariates, satisfies a given parametric model [89]. In this thesis, we will not focus on formal methods for assessing fit in cure rate models. We will instead use graphical methods to assess if the estimated survivor function from a fitted cure rate model approaches an asymptote at the estimated cured proportion for all patients [90]. For models fitted using MCMC, we will use the model diagnostic techniques in Bayesian models which are discussed next to assess fit.

2.10.4 Checking model convergence in Bayesian models

A number of ways can be used to monitor the convergence of the posterior distribution. The Monte Carlo (MC) error, which gives an idea of how precisely the measure of interest is estimated, is given as part of the MCMC output in standard MCMC software such as Bayesian inference Using Gibbs Sampling (BUGS) [91], Just Another Gibbs Sampler (JAGS) [92] and Stan [93]. Small values of the MC error would imply that the resulting estimate is not affected by the MC sampling error and if the chain is ran for long enough, the MC error would decrease to zero. Autocorrelations can also be used to give an idea of convergence with low or high values indicating fast or poor convergence of the Markov Chain. These can be plotted from the MCMC output.

Another way is to use trace plots which show draws from the posterior distribution for each parameter against the number of iterations. Plots with the burn-in period included will show the chains starting at different values and then merging into one stationary

chain indicating the final stationary distribution. The initial iterations can be discarded from these plots if so desired. These plots also show whether the chain of samples is moving quickly around the posterior distribution rather than getting stuck in one region of it.

A more formal way of checking convergence is the Gelman-Rubin diagnostic (GRD) statistic. It compares output from multiple MCMC chains focusing on the variances within and between chains. If the target distribution has been attained by each chain, the posterior variances between and within each chain should be close to each other. These two variances should give a ratio that is close to 1. The square root of this ratio, called the potential scale reduction factor (PSRF) is normally used to give a measure of how close to each other the chains are and therefore whether each of the chains has converged to the target posterior distribution [94, 95]. The GRD is checked for each parameter of interest to ensure each has achieved convergence.

Another formal model checking approach is to use the posterior predictive distribution [96]. This is where, given the observed data D, we can predict future observations $D^{'}$ from the posterior predictive distribution

$$p(D^{'}|D) = \int f(D^{'}|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}. \tag{2.34}$$

This gives the expected distribution of $D^{'}$ having observed the data and averaging over the posterior distribution of $\boldsymbol{\theta}$ given the data. If we have observed data D, the prior distribution $\boldsymbol{\theta}$ shifts to the posterior distribution $p(\boldsymbol{\theta}|D)$, and to a different distribution of variables $D^{'}$ which represent hypothetical or future data. If the replicated data is of the same size and shape as the original D, then we can conclude that the model represents the truth.

2.11 Model selection

Models are usually fitted to assess the effect of one or more covariates on the outcome. Interest is normally on obtaining treatment effects that take into account other factors in the model and thereby avoid confounding. Having fitted models with several covariates, it might be necessary to perform model selection to end up with the simplest model and to avoid over-fitting. In a regression analysis, standard methods for model comparisons can be used to compare nested models for instance the partial likelihood ratio test to assess the significance of a covariate in the model. This test is usually preferred to the Wald and score tests [97]. Where models are not nested information criteria can be used for model selection, for example Akaike and Bayesian Information Criteria which are briefly discussed next.

2.11.1 Choosing between models using the Deviance Information Criterion and other ways of model selection

As with modelling in general, the distribution of the TTE data depends on a vector of unknown parameters we are interested in, $\boldsymbol{\theta}$. In the ML paradigm, model assessment will involve working out the deviance which is the difference in the LL between the full and reduced models [98]. The Akaike's Information Criterion (AIC) is used to compare models fitted using ML based techniques [99] and is given by,

$$AIC = -2\log p(D|\hat{\boldsymbol{\theta}}) + 2p \tag{2.35}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimator of $\boldsymbol{\theta}$ and p is the number of parameters in the model. In the Bayesian setting, the Bayesian Information Criterion (BIC) [100] can also be used in model selection. Another useful measure, the Deviance Information Criterion (DIC) which is made up of a measure of how well a model fits the data (goodness of fit)

and a measure of model complexity, can be used to provide an idea of the effective number of parameters included in the model [101]. In practice, we would use the DIC instead of the BIC as the concern in our case is not to identify the true model but rather to assess the model's predictive ability. Moreover, we would not choose the BIC as it requires specification of the number of parameters while the DIC will estimate the effective number of parameters instead. The DIC is given as

$$DIC = \bar{\boldsymbol{D}} + p_{\boldsymbol{D}} \tag{2.36}$$

where $\bar{\boldsymbol{D}}$ is the posterior expectation of the deviance and is given by

$$\bar{\boldsymbol{D}} = \frac{1}{J'} \sum_{j=1}^{J'} -2 \log p(D|\boldsymbol{\theta}_j) \tag{2.37}$$

where J' is the number of iterations remaining from the MCMC output having discarded the burn-in, and $\log p(D|\boldsymbol{\theta}_j)$ is the LL of the data given the unknown parameters [102] and $p_{\boldsymbol{D}}$ represents the effective number of parameters which is the difference between $\bar{\boldsymbol{D}}$ and the deviance evaluated at the posterior mean of the parameter estimates,

$$p(D|\boldsymbol{\theta}_j) = \bar{\boldsymbol{D}} - \boldsymbol{D}(\bar{\boldsymbol{\theta}}). \tag{2.38}$$

Based on this, the DIC can also be written as

$$\begin{aligned} DIC &= 2\bar{\boldsymbol{D}} - \boldsymbol{D}(\bar{\boldsymbol{\theta}}) \\ &= \boldsymbol{D}(\bar{\boldsymbol{\theta}}) + 2p_{\boldsymbol{D}} \\ &= -2 \log p(D|\boldsymbol{\theta}) + 2p_{\boldsymbol{D}} \end{aligned} \tag{2.39}$$

which bears some resemblance to the AIC. Models with low DIC, AIC or BIC values are considered to be a better fit to the observed data. The use of the DIC for model checking has its own limitations such as $p_{\boldsymbol{D}}$ being invariant to transformation and a

lack of consistency among others. [102]. However, in this work, the focus is on using the DIC to assess practical aspects of the rather complex models fitted by checking for negative values of p_D, for example, which would indicate that the model is not properly specified as well as model selection. In situations where the DIC is not appropriate, the average deviance $\bar{D}$ can, on its own, be used as a measure of fit for complex models by comparing the change in deviance between nested models.

2.12 Summary

In this chapter, we have provided a brief overview of the methods for analysing TTE data in general when interest is on investigation of factors associated with TTR and OS in cancer, and where the possibility of some patients being cured is realistic, how cure rate models can be used to estimate the proportion cured. We have also briefly described how, in the presence of missing or latent variables, the EM algorithm can be used to estimate these outcomes.

The chapter has also briefly discussed how parameters can be estimated using ML based and Bayesian approaches, including checking for model fit using various graphical and more formal means. For the ML approach, we have described how the parameters can be estimated from the score functions and the precision of the estimates can be approximated using the Fisher Information matrix. In the Bayesian approach, we have detailed how the Gibbs sampling and Metropolis-Hastings algorithms can be used to sample from the joint posterior distribution of the unknown parameters, given the observed data. We have finally discussed how to obtain the parameter estimates from the MCMC output, to check for model adequacy and how to select the most appropriate model.

Having described the methods in general, and how they can be applied, the next chapters will focus on how they can be used in practice. The standard methods will be

used to model the TTR and OS patterns using the example data sets. Mixture models will be used to estimate the cured proportion.

Chapter 3

Exploring factors affecting survival outcomes in patients recruited in UK Myeloma and CLL trials

3.1 Introduction

This chapter introduces two datasets that will be used for illustration throughout the thesis. Some of the commonly used TTE methods were applied to the Myeloma IX and CLL datasets to investigate factors associated with TTR and OS following treatment. In this descriptive analysis, the distributions of continuous covariates were briefly explored using means and medians. Cross-tabulations were used to show proportion of patients in groups for categorical variables. Thereafter, K-M curves were used to compare TTR and OS patterns among the groups with accompanying log-rank tests to test differences in the observed patterns.

Following this, Cox PH, Exponential and Weibull regression models were used to model the association of TTR or OS and each covariate in univariable models as an exploratory first step. In the next step, we built a multivariable model using a backward selection approach. Starting with all potential variables in the model, the least

significant covariate at each stage, according to the likelihood ratio test, was dropped. To account for the imbalance likely to result from missing data, models with and without the covariate were fitted to the same number of individuals each time, retaining all available data for that covariate. This process was repeated until no further variables can be dropped without compromising model fit. In the analysis of the two example datasets, backward elimination was performed in such a way that important predictors of interest in this analysis such as log-RD were forced into the final model as we believe they are associated with both TTR and OS after treatment. The backward selection was performed in Cox PH models and the variables in the resulting final models were entered into multivariable Exponential and Weibull TTR and OS models, to enable easy comparison of models. This analysis was based on complete cases because some of the variables had a lot of missing cases. All patients with missing data for any of the variables were excluded from the analysis. Alternative methods for model selection include the forward selection technique where each covariate is evaluated before being added to the model and using the AIC or BIC which were discussed in Section 2.11.1.

To assess validity of the PH assumption, plots of Schoenfeld residuals were used to assess each of the covariates in the univariable and multivariable Cox PH models. We did not perform further assessments of fit in the Exponential and Weibull model as the focus at this point was on the role of covariates on the TTR and OS and not on modelling the hazard function itself.

3.1.1 Outcomes and variable descriptions

The two outcomes TTR and OS were defined as follows:

1. Time to relapse (TTR); this was defined as the time from measurement of the RD percentage to a relapse and/or death from a relapse whichever was sooner. Those who died from causes other than Myeloma or were still alive and disease-free at the end of follow-up were censored at that point.

2. Overall survival (OS); this was defined as the time from measurement of the RD percentage to death from any cause. Those still alive at the end of the follow-up period were censored.

The two example datasets are now briefly described.

3.2 The Myeloma data set

The Myeloma IX trial was a randomised controlled trial of 1,970 newly diagnosed multiple myeloma patients aged 18 and above from 120 centres in the UK between May 2003 and November 2007 [103]. It was a phase III trial to evaluate bisphosphonate and thalidomide - based therapy for myeloma. Bisphosphonates are drugs that help to prevent bone fractures and bone pain in people with myeloma while Thalidomide drugs work to slow the growth of myeloma cells. The aims of this trial were to: 1) Compare new combinations of drugs including thalidomide with treatments already used, 2) Compare two bisphosphonate based treatments, zoledronic acid and clodronate, and 3) Investigate thalidomide as maintenance treatment. The Myeloma IX trial design allowed for testing the effect of treatments in combination and their interactions.

This analysis focused on 1,099 patients who were in the intensive treatment arm of the trial as these patients were considered fit to undergo this kind of treatment. Those in the intensive group had their RD percentage measured at end of treatment. As this was a measurement for research purposes patients were given the choice of providing the measurement or not. As a result, there were no RD percentage measurements recorded for 661 patients in this group. A comparison of those who did and did not have a recorded RD percentage showed that age at randomisation was slightly higher (mean age (SD) 58.3 (0.29) years), for missing cases, compared to those with available data (56.8 (0.34) years, t-test p-value = 0:001). Moreover, 22.8% (151/661 patients) of those with missing RD percentage data were 65 years or older, compared to only

14.6% (64/438 patients) for those with complete data. This suggests that older patients were less likely to agree to the invasive procedure of measuring the RD from the bone marrow, which may affect the generalisability of results. Finally, 11 patients were excluded from the analysis because their relapse or death was recorded at or before RD assessment, leaving 427 patients in the final data set, Figure 3.1. As the focus of this thesis was on the association between RD percentage measurement and subsequent clinical events, the date of measurement of RD percentage was treated as the baseline or time zero. Patients missing this measure were excluded from the analysis.

3.2.1 Predictors of TTE outcomes in Myeloma

The RD percentage, and in most cases its natural logarithm (log-RD), was the main explanatory variable of TTR and OS that was considered in this analysis. The TTR and OS were calculated for each patient using the date the RD percentage was measured as detailed Section 3.1.1. The RD percentage measurements were on the continuous scale, however some patients (40/427) did not have exact values and for these, cut-offs were used instead. For example $< 0.01\%$ was used to represent an RD percentage of less than 0.01%. These were represented by specific values such as 0.01 ignoring the inequality. Other covariates used in the analyses were: sex, age at randomisation, stage (I, II, III), a measure of the paraprotein levels in grams per litre (g/L) in the blood, albumin in grams per decilitre (g/dL) and beta2-microglobulin in milligrams per litre (mg/L) at baseline. For brevity, paraprotein (g/L), albumin (g/dL) and beta2-microglobulin (mg/L) will in some instances just be referred to as paraprotein, albumin and beta2 respectively. Stage is a marker of disease severity at baseline while paraprotein, albumin and beta2 are hypothesised to be related to the tumour re-growth rate after treatment. The data also included cytogenetic markers: hyperdiploidy, and the translocations t(4;14), t(11;14), t(14;16), t(6;14) and t(14;20), as well as dp53 and gain(1q21) [104], which have already been investigated for their association with TTR

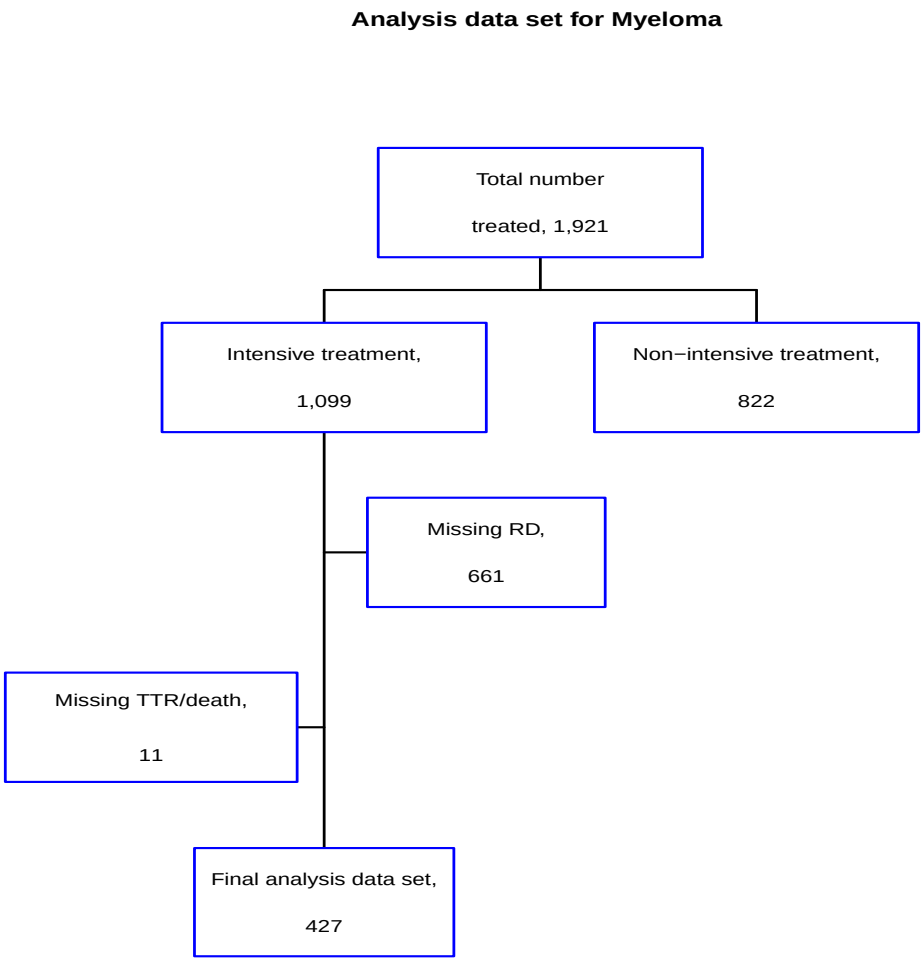


Figure 3.1: Derivation of the Myeloma IX final data set

and OS in Myeloma. Treatment is normally included when analysing survival outcomes when analysing RCTs. As this work was focused on what happens following treatment, it was not included in the main analyses in analysing the Myeloma data. This was after quickly checking that the randomisation into four treatment groups was balanced and that there were no obvious differences in outcomes among treatment arms (data not

shown). When we included treatment in some early analyses, it did not have a significant effect itself on the TTR or OS and it did not change other effect of the other observed outcome predictors. In this thesis, the TTE was calculated from the date of the end of treatment to the date of relapse, death or censoring. As there was a time-lag between recruitment and measurement of RD percentage after treatment - the baseline in this analysis, all covariates were assumed to have the same values at time zero as on the date of recruitment to the trial. We further assumed that treatment had no effect after it ceased.

As in some studies, there were some missing values for some of the covariates in the Myeloma dataset after having excluded all those without RD measurement as shown in Table 3.1. Complete information was collected for log-RD, age, and sex, while 1 patient did not have a recorded albumin value. Other than that, the binary covariates; hyperdiploidy, t(4;14), t(11;14), t(14;16), t(6;14), dp53 and gain(1q21) had over 40% missing information. Stage is calculated from beta2 and albumin, therefore, 117 (27.4%) did not have data for both stage and albumin. This illustrates that there was a potential problem with missing data in this analysis. The analysis first focused on the available data in a complete case (CC) analysis. Since using only the available data might lead biases in the estimates, ways of handling missing data were then applied to the data and are detailed in Section 3.4.

3.2.2 Descriptive summary of the categorical variables in Myeloma

Cross-tabulations of the categorical variables and the two outcomes TTR and OS are shown in Table 3.2. As the length of follow-up to relapse or death varied between patients, it is inappropriate to compare those with and without each event using these crude summaries; we comment only on the overall proportions in the different categories (final column in the table). Information on sex was available on all the 427 patients in the final data set, males constituted 62.8% of the total. Disease stage was

Covariate	Covariate type	Available (%)	Missing (%)
Log-RD	Continuous	427 (100)	0 (0)
Age	Continuous	427 (100)	0 (0)
Sex	Binary	427 (100)	0 (0)
Stage	Categorical	310 (72.6)	117 (27.4)
Hyeperdiploidy	Binary	246 (57.6)	181 (42.4)
t(4;14)	Binary	249 (58.3)	178 (41.7)
t(11;14)	Binary	247 (57.8)	180 (42.2)
t(14;16)	Binary	247 (57.8)	180 (42.2)
t(6;14)	Binary	243 (56.9)	184 (43.1)
t(14;20)	Binary	245 (57.4)	182 (42.6)
dp53	Binary	240 (56.2)	187 (43.8)
Gain(1q21)	Binary	222 (52.0)	205 (48.0)
Paraprotein	Continuous	334 (78.2)	93 (21.8)
Beta2	Continuous	310 (72.6)	117 (27.4)
Albumin	Continuous	426 (99.8)	1 (0.2)

Table 3.1: A summary of the covariates used in the analysis of the Myeloma dataset showing proportion missing for each covariate. The total number of patients in the analysis was 427

ascertained for 310 patients with roughly a third in each of the stages I, II and III. Over half (53.2%) of the 246 patients tested positive for hyperdiploidy. Just over 14% of 249 and 16% of 247 patients were positive for t(4;14) and t(11;14) translocations respectively while the proportions of those with t(14;16), t(6;14), t(14;20) and dp53 were even lower. Finally, 36% of 222 patients had positive gain(1q21). The cross-tabulations also showed that there were cells with frequencies less than 5 for some of the translocations which precludes the use of these variables in our statistical models [105]. These variables were thus not included in subsequent analyses.

3.2.3 Overall TTR and OS patterns in Myeloma

We next graphically looked at the TTR and OS from the time of RD percentage measurement for these patients by plotting K-M survival curves, Figure 3.2. The median relapse time for the Myeloma patients after RD measurement was 2 years inter-quartile range (IQR) [1.7, 2.3 years]. Most of the relapses occurred within the first 2 years as evidenced by the steep curve. After 5 years, the curve seemed to flatten out at around 20%, suggestive of a plateau with heavy censoring up to the end of follow-up (only 4 events were observed in 51 patients who were followed up for 5 years or longer). For the OS, the median time to death after RD measurement was more than 6 years. Unlike in the TTR curve, deaths continued to occur evenly throughout the follow-up time.

3.2.4 Comparing TTR patterns by categorical variables in Myeloma

Comparisons of TTR patterns were made based on stage, sex, hyperdiploidy, t(4;14), t(11;14) and gain(1q21), Figure 3.3. Log-rank tests were used to test for statistical significance among groups in the observed TTR patterns. There were no obvious differences in the TTR patterns among males and females. Patients with Stage I disease had longer TTR compared to Stages II and III although these differences not quite significant at the 5% level. Those without hyperdiploidy had a shorter median TTR (1.6

	TTR		OS		
Variable (n)	Not relapsed (%)	Relapsed (%)	Alive (%)	Died (%)	Total (%)
Sex (427)					
Male	67 (63.8)	201 (62.4)	161 (62.9)	107 (62.6)	268 (62.8)
Female	38 (36.2)	121 (37.6)	95 (37.1)	64 (37.4)	159 (37.2)
Stage (310)					
I	32 (43.8)	74 (31.2)	72 (38.7)	34 (27.4)	106 (34.2)
II	22 (30.2)	91 (38.4)	73 (64.6)	40 (32.3)	113 (36.4)
III	19 (26.0)	72 (30.4)	41 (22.0)	50 (40.3)	91 (29.4)
Hyperdip. (246)					
No	23 (42.6)	92 (47.9)	58 (38.4)	57 (60.0)	115 (46.8)
Yes	31 (57.4)	100 (52.1)	93 (61.6)	38 (40.0)	131 (53.2)
t(4;14) (249)					
No	53 (98.2)	161 (82.6)	138 (92.0)	76 (76.8)	214 (85.9)
Yes	1 (1.8)	34 (17.4)	12 (8.0)	23 (23.2)	35 (14.1)
t(11;14) (247)					
No	43 (81.1)	164 (84.5)	122 (81.9)	85 (86.7)	207 (83.8)
Yes	10 (18.9)	30 (15.5)	27 (18.1)	13 (13.3)	40 (16.2)
t(14;16) (247)					
No	52 (98.1)	188 (96.9)	147 (98.0)	93 (95.9)	240 (97.2)
Yes	1 (1.9)	6 (3.1)	3 (2.0)	4 (4.1)	7 (2.8)
t(6;14) (243)					
No	50 (98.0)	191 (99.5)	144 (98.6)	97 (100.0)	241 (99.2)
Yes	1 (2.0)	1 (0.5)	2 (1.4)	0 (0.0)	2 (0.8)
t(14;20) (245)					
No	53 (100.0)	188 (97.9)	148 (99.3)	93 (96.9)	241 (98.4)
Yes	0 (0.0)	4 (2.1)	1 (0.7)	3 (3.1)	4 (1.6)
dp53 (249)					
No	49 (100.0)	179 (93.7)	144 (99.3)	84 (88.4)	228 (95.0)
Yes	0 (0.0)	12 (6.3)	1 (0.7)	11 (11.6)	12 (5.0)
Gain(1q21) (222)					
No	37 (77.1)	105 (60.3)	98 (73.1)	44 (50.0)	142 (64.0)
Yes	11 (22.9)	69 (39.7)	36 (26.9)	44 (50.0)	80 (36.0)

Table 3.2: Summary of categorical variables in the Myeloma dataset by TTR and OS outcomes. The number of available observations for each variable are shown in parentheses next to the variable name.

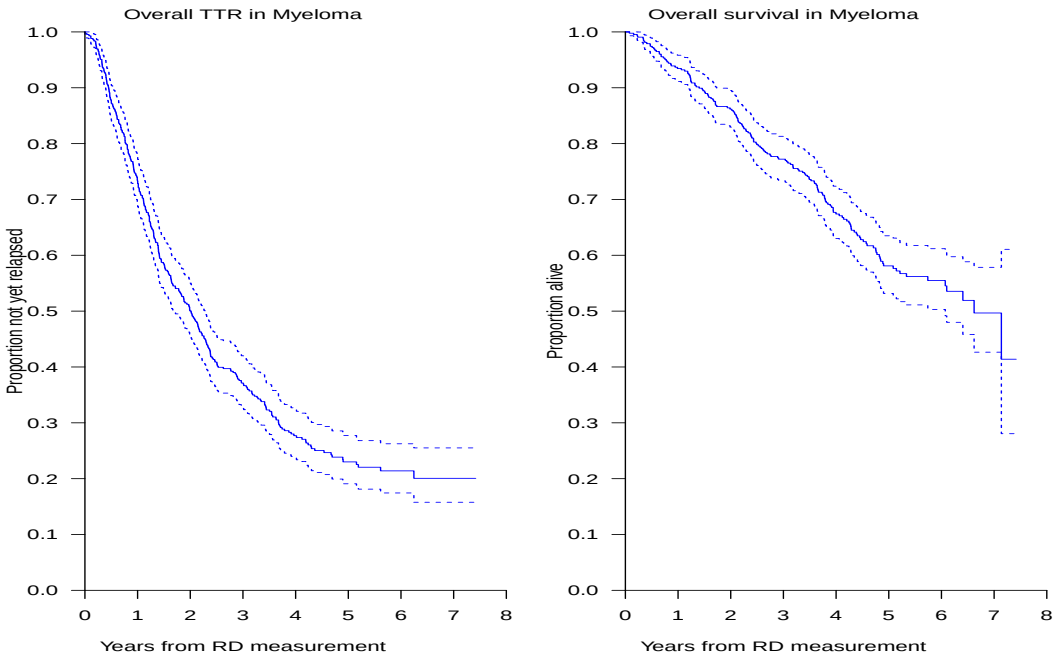


Figure 3.2: K-M estimates of the overall TTR (left) and OS (right) in the Myeloma dataset. Dotted lines represent 95% CI limits.

years) compared to those with hyperdiploidy (2.2 years), $p < 0.081$ based on the log-rank test. Those with the t(4;14) translocation had a median TTR of just over 6 months, while those without it relapsed after 2.2 years on average, log-rank $p < 0.001$. Moreover, all but one of the patients with t(4:14) eventually relapsed by the end of follow-up. There were no differences in the relapse patterns by t(11;14), log-rank $p = 0.971$. Those without gain(1q21) had a longer median TTR 2.3 years compared to the those with gain(1q21) who had a median TTR of 1.1 years, log-rank $p < 0.001$.

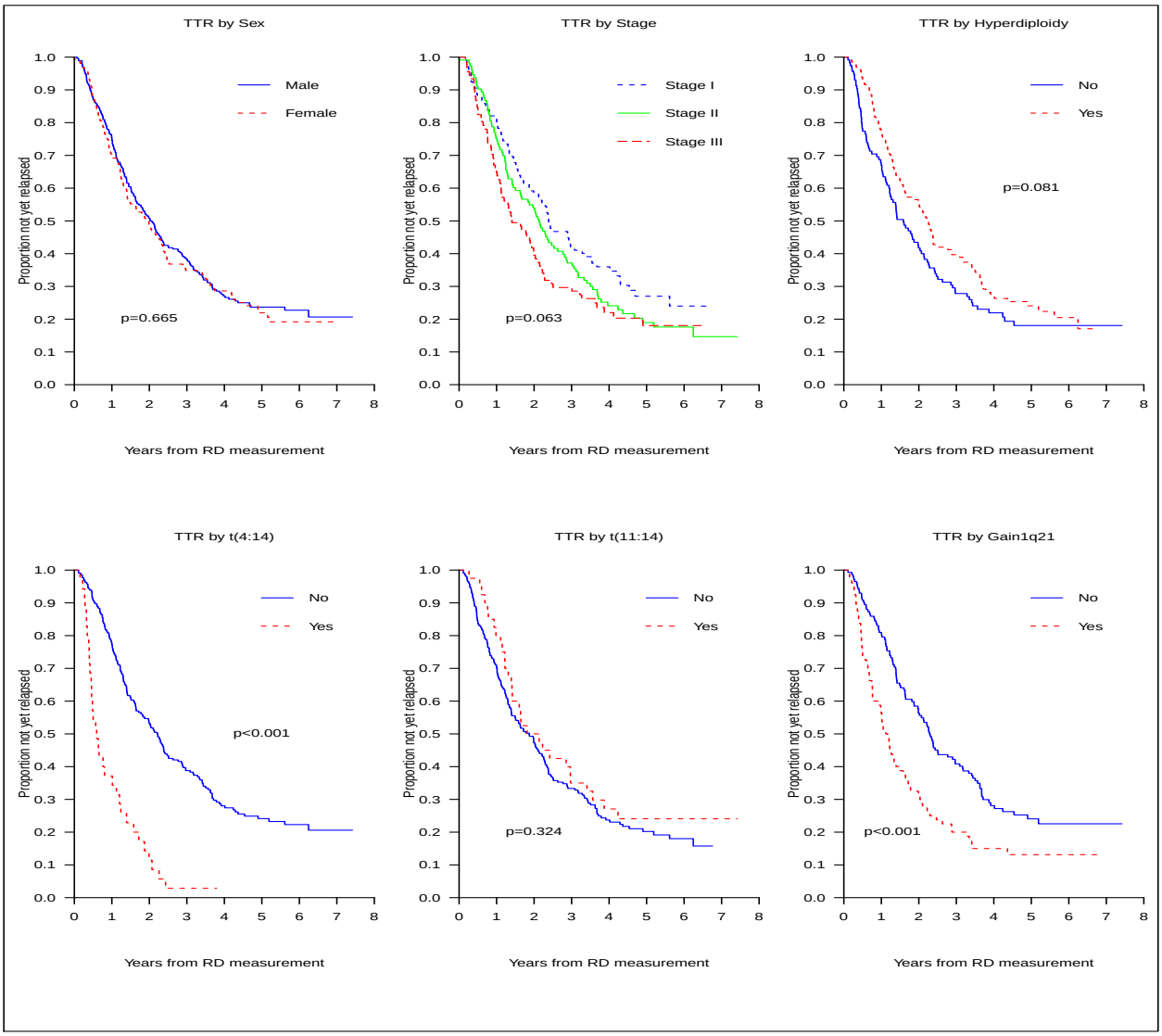


Figure 3.3: K-M estimates of the overall TTR and OS patterns in the Myeloma dataset

3.2.5 Comparing observed OS patterns by categorical variables in Myeloma

There were differences in OS patterns by disease stage (visibly between stage I or II and stage III), log-rank $p < 0.001$, Figure 3.4. Those without hyperdiploidy had poorer OS while having t(4;14) and gain(1q21) were associated with a shorter time to death.

There were no differences in OS between males and females and between those with and without t(11;14).

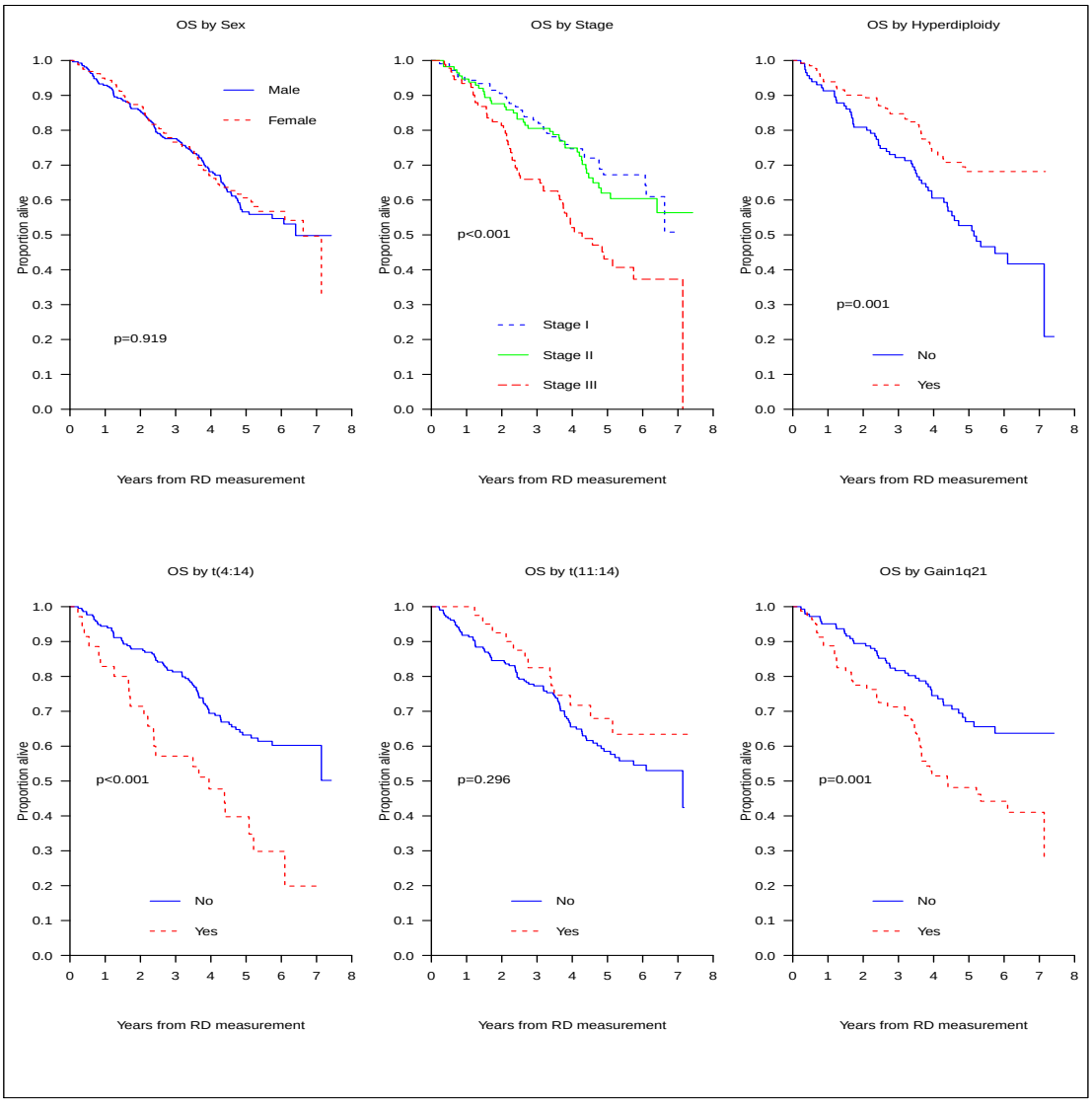


Figure 3.4: K-M estimates of OS by categorical variables in the Myeloma dataset

		Original scale		Z-standardised	
	n	Mean (SD)	Median [IQR]	Mean (SD)	Median [IQR]
Age (years)*	427	56.8 (7.25)	57 [52, 62]	-	-
Log-RD*	427	-0.72 (0.54)	-0.70 [-1.00, -0.40]	-	-
Paraprotein (g/L)	334	4.07 (6.47)	0.00 [0.00, 6.68]	1.29×10^{-8} (1.00)	-0.63 [-0.63, 0.40]
Beta2 (mg/L)	310	4.98 (4.12)	3.50 [2.60, 5.90]	-1.72×10^{-9} (1.00)	-0.36 [-0.58, 0.22]
Albumin (g/dL)	426	3.55 (0.73)	3.60 [3.10, 4.10]	-4.90×10^{-10} (1.00)	0.06 [-0.62, 0.75]

Table 3.3: Summary by continuous variables in the Myeloma dataset. Age and log-RD, which was already log-transformed, were not z-standardised (*). The summaries for paraprotein, beta2 and albumin were based on 334, 310 and 426 patients respectively.

3.2.6 Descriptive summaries of the continuous variables in Myeloma

Age, log-RD, paraprotein, beta2 and albumin were the continuous variables in the Myeloma dataset. The means, standard deviations (SD), medians and IQRs for these continuous variables are shown in Table 3.3. For each covariate, the number of observations is included recognising that some patients had missing data as reported in Table 3.1 had $n < 427$. The median age at randomisation [IQR] was 57 [52, 62] years. The mean (SD) log-RD was -0.72 (0.54) at the end of treatment. For paraprotein, the mean (SD) level in the blood was 4.07 (6.47) at baseline. For beta2 and albumin, the means (SDs) were 4.98 (4.12) and 3.55 (0.73) respectively. Looking at the medians, it was clear that paraprotein and beta2 were positively skewed, with around half of the patients having a paraprotein measure of zero. The values of paraprotein, beta2 and albumin were z-standardised in order to model the effect of these covariates as increases or decreases in the hazard function for each SD increase/decrease in the value of the variable when used in TTE model thereby making interpretation easier.

Following the exploratory analysis, univariable Cox PH, Exponential and Weibull regression models were fitted to the Myeloma data to investigate the association of the various covariates with TTR and OS. In these models, RD percentage was fitted on the

log-scale (log-RD), while z-standardised values of paraprotien, beta2 and albumin were used. The ML methods, discussed in Chapter 2, were used to estimate parameters from these models in Stata, first with each variable fitted singly in univariable models. Variables found to be significantly associated with either outcome ($p < 0.1$) were then included in a multivariable analysis together with log-RD and the z-standardised paraprotein, beta2 and albumin. These four variables were included in all multivariable models because log-RD was the main predictor of TTE outcomes following treatment in this thesis, alongside the tumour growth rate whose markers are paraprotein, beta2 and albumin. Because stage is derived from beta2 and albumin, stage was not included in the multivariable models with these two variables used instead. These three covariates will be used in modelling the role of the latent variable tumour growth rate in a structural equation modelling approach in Chapter 6. Model selection using backward elimination was performed as described earlier in this chapter. We report estimates of the log-hazard ratios (Log-HR) in univariable models and adjusted log-hazard ratios (Adj. Log-HR) for all models as well as estimates of the shape parameter (γ) from the Weibull models.

3.2.7 Modelling independent factors associated with TTR in Myeloma

Results from Cox PH, Exponential and Weibull models for factors associated with TTR following treatment are shown in Table 3.4. Each log-percentage increase in the RD was associated with a significant, 0.214 (0.008, 0.421), increase in the log-hazard of a relapse based on the Cox PH model. This effect was similar to that from the Exponential 0.200 (-0.006, 0.405) and Weibull 0.200 (-0.005, 0.406) model estimates, although these were non-significant. Age and sex were not independently associated with the TTR in any of the univariable models (results not shown). They were thus not considered in the multivariable analyses. With respect to the disease stage at recruitment, the log-hazard was 0.221 (-0.086, 0.528) times higher in stage II patients when com-

pared to those in stage I, although this was not significant, while that of stage III patients was 0.387 (0.062, 0.712) times significantly higher than that of stage I patients based on the Cox PH model. Estimates of the log-HR for stage from the Exponential and Weibull models were generally similar to those from the Cox PH model. Each SD increase in the paraprotein level was significantly associated with a 0.203 (0.096, 0.311) increase in the log-hazard based on the Cox PH model, a 0.207 (0.098, 0.316) increase in the log-hazard based on the Exponential and a 0.215 (0.106, 0.324) increase in the log-hazard from the Weibull model. Similarly, higher beta2 levels were significantly associated with an increase in the log-HR of 0.193, 0.187, and 0.192 for each SD increase in beta2 in the Cox PH, Exponential and Weibull models respectively. Contrasting with paraprotein and beta2, each SD increase in albumin was significantly associated with a 0.138, 0.124 and 0.124 decrease in the log-HR based on the Cox PH, Exponential and Weibull models respectively. Hyperdiploidy and t(11;14) were not significant associated with the TTR among these patients. On the other hand, t(4;14) and gain(1q21) were significantly associated with shorter TTR.

The estimates from the univariable models were generally similar for all three survival models. In all univariable models for TTR, the 95% CI for the shape parameter in the Weibull models included 1. As the Exponential distribution is a special case of the Weibull with shape = 1, there is little to gain from fitting the Weibull model for TTR over the Exponential model, which assumes constant hazard. Among the three survival models however, the Cox PH model would be the model of choice where interest is on the role of covariates on the TTR as it not vulnerable to the misspecification of the baseline hazard. We thus only assessed whether it was reasonable to assume PH in the Cox PH model.

		Cox PH	Exponential	Weibull
Parameter	n	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Intercept	427	-	-1.043 (-1.219, -0.867)	-1.068 (-1.276, -0.861)
Log-RD		0.214 (0.008, 0.421)	0.200 (-0.006, 0.405)	0.200 (-0.005, 0.406)
Shape (γ)		-	-	1.022 (0.932, 1.120)
Intercept	310		-1.358 (-1.586, -1.130)	-1.440 (-1.709, -1.172)
Stage II		0.221 (-0.086, 0.528)	0.210 (-0.096, 0.517)	0.213 (-0.093, 0.520)
Stage III		0.387 (0.062, 0.712)	0.376 (0.051, 0.700)	0.387 (0.062, 0.712)
Shape (γ)		-	-	1.066 (0.956, 1.185)
Intercept		-	-1.178 (-1.301, -1.054)	-1.270 (-1.451, -1.090)
Paraprotein	334	0.203 (0.096, 0.311)	0.207 (0.098, 0.316)	0.215 (0.106, 0.324)
Shape (γ)		-	-	1.080 (0.974, 1.197)
Intercept	310	-	-1.168 (-1.295, -1.040)	-1.249 (-1.434, -1.064)
Beta2		0.193 (0.076, 0.310)	0.187 (0.071, 0.303)	0.192 (0.076, 0.308)
Shape (γ)		-	-	1.069 (0.962, 1.188)
Intercept	426	-	-1.184 (-1.294, -1.075)	-1.210 (-1.366, -1.054)
Albumin		-0.138 (-0.241, -0.035)	-0.124 (-0.225, -0.023)	-0.124 (-0.225, -0.023)
Shape (γ)		-	-	1.022 (0.932, 1.120)
Intercept	246	-	-1.006 (-1.211, -0.802)	-1.048 (-1.295, -0.800)
Hyperdiploidy: Yes		-0.252 (-0.535, 0.032)	-0.230 (-0.513, 0.053)	-0.232 (-0.515, 0.051)
Shape (γ)				1.036 (0.922, 1.165)
Intercept	249	-	-1.251 (-1.406, -1.097)	-1.394 (-1.619, -1.170)
t(4;14): Yes		1.284 (0.899, 1.668)	1.263 (0.893, 1.633)	1.373 (0.984, 1.763)
Shape (γ)		-	-	1.115 (0.994, 1.251)
Intercept	247	-	-1.073 (-1.226, -0.920)	-1.112 (-1.318, -0.905)
t(11;14): Yes		-0.196 (-0.585, 0.194)	-0.246 (-0.636, 0.143)	-0.252 (-0.641, 0.138)
Shape		-	-	1.033 (0.920, 1.161)
Intercept	222	-	-1.289 (-1.481, -1.098)	-1.345 (-1.592, -1.098)
Gain(1q21): Yes		0.590 (0.284, 0.895)	0.560 (0.256, 0.863)	0.570 (0.265, 0.875)
Shape (γ)		-	-	1.045 (0.926, 1.180)

Table 3.4: Log-hazard ratio estimates for each covariate from univariable Cox PH, Exponential and Weibull models for TTR applied to the Myeloma dataset in a CC analysis (details of n for each covariate are shown based on the summary in Table 3.1). Estimates of the intercept for the Exponential and Weibull models as well as the shape parameter (γ) for the Weibull model are also reported.

Test for the PH assumption in univariable models for TTR in Myeloma

To test whether the PH holds in the individual univariable Cox PH models, plots of the scaled Schoenfeld residuals against time for all the covariates in the univariable models except age and sex were used, Figure 3.5. Alongside the plots, the proportionality assumption for each independent predictor was tested using the stphtest command in Stata. A p-value of greater than 0.05 from this test for a given covariate would imply that there is no strong evidence to suggest violation of proportionality and we therefore do not reject the PH assumption in that model.

From the plots the red horizontal line around 0 suggested that the PH generally held for most of the covariates used in the univariable Cox PH models for TTR in the Myeloma dataset. Further, the p-values from the test for the PH assumption obtained in Stata were > 0.05 for all covariates except hyperdiploidy (labelled hyperd in this figure), ($p = 0.025$) and gain(1q21) whose p-value was 0.017. There appeared to be deviations in the scaled residuals after four years for the two covariates.

3.2.8 Modelling independent factors associated with OS in Myeloma

The factors associated with OS were assessed using Cox PH, Exponential and Weibull models, Table 3.5. Each unit increase in the log-RD was associated 0.087 (-0.203, 0.377) increase in the log-hazard in the Cox PH model, a 0.085 (-0.203, 0.374) increase in the log-hazard in the Exponential model and a 0.076 (-0.214, 0.366) increase in the log-hazard in the Weibull model, respectively. However, unlike in the TTR models, this effect was not significant in any of the three models. Age and sex were again not associated with the OS in all models (results not shown). As with TTR, an advanced disease stage (stage II and stage III) was predictive of poorer OS when compared to those who had stage I disease. However, the log-HR for those with stage II disease was not significant when compared to those with stage I disease in all three models. On the

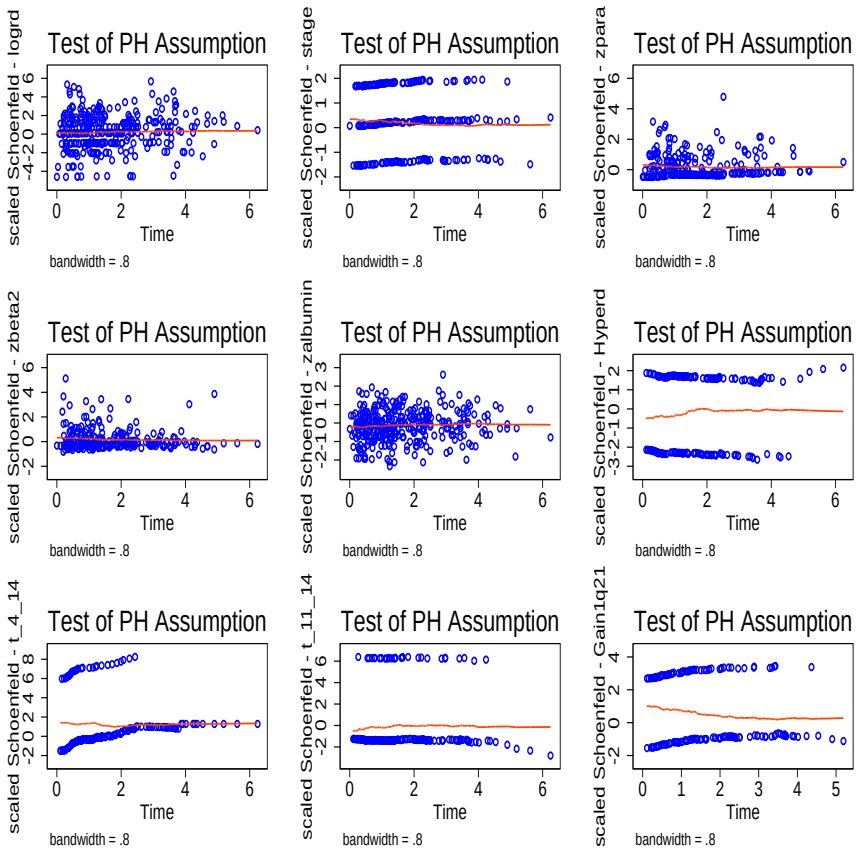


Figure 3.5: Scaled Schoenfeld residuals plotted against time from the univariable Cox PH models for TTR in the Myeloma dataset ignoring missing data

other hand, stage III patients had a significantly higher log-HR 0.758 (0.321, 1.196) in the Cox PH, 0.725 (0.289, 1.160) in the Exponential and 0.775 (0.338, 1.212) in the Weibull models respectively.

Paraprotein was associated with a slight, but non-significant effect on the OS in all three models. Each SD increase in beta2 associated with a significant 0.278 (0.145, 0.410), 0.266 (0.133, 0.399) and 0.282 (0.149, 0.414) increase in the log-HR in the Cox PH, Exponential and Weibull models respectively. Each SD increase in albumin levels was on the other hand significantly associated with a lower hazard of death or better

OS. Having hyperdiploidy was predictive of better OS while the markers t(4;14) and gain(1q21) were both significantly associated with poor OS in all three models. There were no differences in OS among those with and without t(11;14).

Once again, the estimates of the log-HR were similar across the different models just as in the TTR models. In contrast to the TTR models, the shape parameter in the Weibull model for OS was significantly > 1 for all models as the 95% CI did not include a 1, indicating the hazard of death was not constant over time and that perhaps Weibull model would be a better model for OS than the Exponential model. However, on comparing the AICs for log-RD, stage, paraprotein, beta2 and albumin which were respectively (871.3, 609.5, 656.2, 608.5, and 866.1) from the Weibull models and (879.7, 619.5, 666.5, 618.4, 874.8) from the Exponential models, it would seem the Weibull was the better model for OS for these particular covariates. The Cox PH model would still be considered ideal for modelling the OS because of its non-parametric handling of the baseline hazard.

Test for the PH assumption in univariable models for OS in Myeloma

Again the scaled Schoenfeld residuals were used to test the PH assumption in the univariable Cox PH models for OS, Figure 3.6. Alongside the plots, the proportionality assumption was formally tested for each covariate. Residuals from all univariable Cox PH models for OS suggested that there was no significant departure from the PH assumption. Moreover, all p-values from the global test for the PH assumption for each covariate were > 0.05. We thus conclude that it was reasonable to model the OS using the Cox PH models in Myeloma.

		Cox PH	Exponential	Weibull
Parameter	n	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Intercept	427	-	-2.257(-2.504, -2.011)	-2.671 (-3.039, -2.303)
Log-RD		0.087 (-0.203, 0.377)	0.085 (-0.203, 0.374)	0.076 (-0.214, 0.366)
Shape (γ)		-	-	1.267 (1.104, 1.453)
Intercept	310	-	-2.602 (-2.938, -2.266)	-3.161 (-3.642, -2.680)
Stage II		0.075 (-0.384, 0.534)	0.096 (-0.361, 0.554)	0.093 (-0.364, 0.550)
Stage III		0.758 (0.321, 1.196)	0.725 (0.289, 1.160)	0.775 (0.338, 1.212)
Shape (γ)				1.356 (1.156, 1.590)
Intercept	334	-	-2.359 (-2.531, -2.187)	-2.892 (-3.263, -2.520)
Paraprotein		0.093 (-0.066, 0.251)	0.083 (-0.075, 0.242)	0.091 (-0.067, 0.249)
Shape (γ)		-	-	1.348 (1.151, 1.579)
Intercept	310	-	-2.347 (-2.526, -2.168)	-2.877 (-3.255, -2.499)
Beta2		0.278 (0.145, 0.410)	0.266 (0.133, 0.399)	0.282 (0.149, 0.414)
Shape (γ)		-	-	1.346 (1.148, 1.577)
Intercept	426	-	-2.320 (-2.470, -2.169)	-2.732 (-3.040, -2.425)
Albumin		-0.156 (-0.298, -0.014)	-0.150 (-0.293, -0.008)	-0.153 (-0.294, -0.012)
Shape (γ)		-	-	1.271 (1.108, 1.457)
Intercept	246	-	-2.064 (-2.324, -1.805)	-2.405 (-2.836, -1.973)
Hyperdiploidy: Yes		-0.673 (-1.084, -0.262)	-0.656 (-1.067, -0.246)	-0.672 (-1.083, -0.261)
Shape (γ)				1.225 (1.019, 1.472)
Intercept	249	-	-2.491 (-2.716, -2.266)	-2.864 (-3.281, -2.447)
t(4;14): Yes		0.890 (0.421, 1.357)	0.875 (0.409, 1.342)	0.912 (0.445, 1.380)
Shape (γ)				1.238 (1.035, 1.481)
Intercept	247	-	-2.304 (-2.516, -2.091)	-2.665 (-3.075, -2.256)
t(11;14): Yes		-0.310 (-0.894, 0.274)	-0.315 (-0.899, 0.268)	-0.322 (-0.906, 0.262)
Shape (γ)				1.234 (1.029, 1.479)
Intercept	222	-	-2.641 (-2.937, -2.346)	-2.992 (-3.464, -2.519)
Gain(1q21): Yes		0.707 (0.288, 1.126)	0.699 (0.281, 1.117)	0.712 (0.294, 1.130)
Shape (γ)				1.223 (1.011, 1.479)

Table 3.5: Log-hazard ratio estimates for each covariate from univariable Cox PH, Exponential and Weibull models for OS applied to the Myeloma dataset in a CC analysis (details of n for each covariate are shown based on the summary in Table 3.1). Estimates of the intercept for the Exponential and Weibull models as well as the shape parameter (γ) for the Weibull model are also reported.

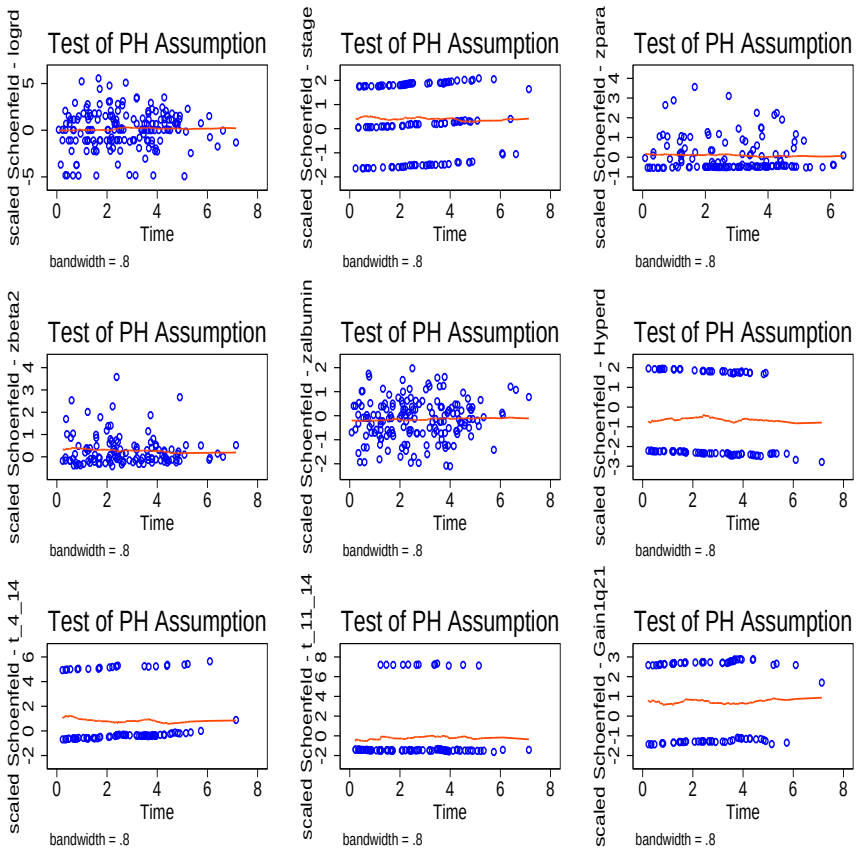


Figure 3.6: Scaled Schoenfeld residuals plotted against time from the univariable Cox PH models for OS in the Myeloma dataset ignoring missing data

3.2.9 Multivariable models for TTR and OS in Myeloma

Having fit univariable models, the aim is naturally, to investigate whether the observed associations are independent and to adjust for potential confounders [106]. We first fit multivariable Cox PH models for TTR and OS, starting with the full model including all the variables used in the univariable models except age, sex and stage. Starting with the full model with all covariates, an a backward model selection procedure as described previously was performed in Stata to come up with the final model. Using the covariates in the final model obtained from the model selection, multivariable Ex-

ponential and Weibull models were again fitted. The results from the final multivariable Cox PH, Exponential and Weibull TTR and OS models are shown in Table 3.6. Due to missing data, only 125 individuals were included in the final multivariable models for TTR and OS in a CC analysis. To enable comparisons between the adjusted and unadjusted estimates, we fit univariable Cox PH models for TTR and OS to data from the 125 patients in the final model, Table 3.7. We also report AICs for each multivariable model.

The adjusted HR estimate in the TTR model was 0.141 (-0.335, 0.616) in the Cox PH model implying an increase in the log-hazard for each unit increase on the log-scale in the RD percentage when paraprotein, beta2, albumin, t(4:14) and gain(1q21) were taken into account. However, this effect was not significant. In comparison, the unadjusted log-HR estimate from the univariable Cox PH model was much higher and significant when the univariable Cox PH model was fitted to exactly the same number of patients, 0.543 (0.135, 0.951), as shown in Table 3.7. Higher levels of paraprotein were predictive of shorter TTRs and the estimate of the log-HR estimate of 0.408 in the univariable model increased to 0.454 when the other covariates were taken into account. Beta2 did not have a significant effect on the TTR in both univariable and multivariable models for this subset of patients. With respect to albumin, the same effect seen in the univariable model was also evident in the multivariable models with each SD increase in albumin being associated with a decrease in the risk of relapsing following treatment. Finally, t(4:14) and gain(1q21) were predictive of poorer outcomes in both the univariable and multivariable models.

For OS, the adjusted log-HR (0.321) for log-RD was almost half of that in the univariable Cox PH model (0.634). Moreover, the effect in the multivariable model was no longer significant. Each SD increase in the paraprotein level was significantly associated with an even higher increase in the hazard in the multivariable Cox PH model compared to the univariable model. Again beta2 was not associated with the time to

death in both models, while the effect of albumin and gain(1q21) remained significantly associated with OS.

Estimates from the Exponential and Weibull models were generally similar in all multivariable models. The AIC was lowest in the Weibull model suggesting that it was the better model. However, because of its flexibility, we may still want model these outcomes using the Cox PH model. Plots of the scaled Schoenfeld residuals from the multivariable models showed that the PH assumption was not violated, Figure E.1 and Figure E.2 in Appendix E. By only including patients with available data on the covariates in the final models for both TTR and OS in the Myeloma dataset, we discarded data from more than half of patients in the trial which could lead to biased results due to missing data. We thus performed some sensitivity analyses by discarding covariates with the highest proportion of missing data to compare estimates.

3.2.10 Sensitivity analysis

By considering only the CC analysis, the final model had 125 patients. This was a drop of 302 patients and differences in the number of cases in each model could have led to the major differences between the univariable and multivariable model results. From the summary of available data in Table 3.1, it was clear that the variables t(4:14) and gain(1q21) had data from only 249 and 222 patients respectively, which had a large impact on the number of missing cases overall. For a more appropriate comparison we fit univariable and multivariable Cox PH models for both TTR and OS with only log-RD, paraprotein, beta2 and albumin, ignoring t(4:14) and gain(1q21), to see if the estimates from these models would be different. The resulting dataset had complete information from 243 individuals. Since all four variables were included in the previous multivariable models we did not perform model selection. The results are shown in Table 3.8.

The effect of log-RD on the TTR was significant in the univariable model 0.352 (0.066, 0.638) but not in the adjusted model taking into account paraprotein, beta2

	Cox PH	**Exponential**	**Weibull**
Parameter	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
TTR model			
Intercept	-	-1.421 (-1.827, -1.015)	-1.924 (-2.441, -1.407)
Adjusted log-HRs			
Log-RD	0.141 (-0.335, 0.616)	0.128 (-0.342, 0.597)	0.147 (-0.339, 0.633)
Paraprotein	0.454 (0.221, 0.688)	0.344 (0.121, 0.566)	0.428 (0.196, 0.660)
Beta2	0.048 (-0.194 0.289)	0.038 (-0.189, 0.265)	0.037(-0.202, 0.276)
Albumin	-0.301 (-0.560, -0.042)	-0.272 (-0.520, -0.024)	-0.333 (-0.589, -0.077)
t(4;14): Yes	1.460 (0.916, 2.004)	1.095 (0.585, 1.605)	1.462 (0.910, 2.015)
Gain(1q21): Yes	0.702 (0.277, 1.127)	0.512 (0.096, 0.929)	0.659 (0.237, 1.081)
Shape (γ)	-	-	1.367 (1.163, 1.605)
AIC	771.1	333.8	323.4
OS model			
Intercept	-	-2.643 (-3.211, -2.076)	-3.212 (-4.012, -2.413)
Adjusted log-HRs			
Log-RD	0.321 (-0.344, 0.987)	0.297 (-0.361, 0.954)	0.298 (-0.366, 0.963)
Paraprotein	0.362 (0.067, 0.658)	0.331 (0.042, 0.620)	0.385 (0.091, 0.679)
Beta2	0.185 (-0.044, 0.413)	0.178 (-0.052, 0.407)	0.190 (-0.041, 0.420)
Albumin	-0.301 (-0.633, 0.031)	-0.297 (-0.629, 0.035)	-0.309 (-0.642, 0.025)
Gain(1q21): Yes	0.720 (0.134, 1.306)	0.651 (0.073, 1.230)	0.709 (0.124, 1.294)
Shape (γ)	-	-	1.346 (1.046, 1.732)
AIC	416.9	243.1	240.3

Table 3.6: Multivariable Cox PH, Exponential and Weibull models for TTR and OS in the Myeloma dataset. The multivariable TTR and OS models were fitted to data from $n = 125$ patients

	Cox PH for TTR	**Cox PH for OS**
	Log-HR (95% CI)	Log-HR (95% CI)
Log-RD	0.543 (0.135, 0.951)	0.634 (0.069, 1.198)
Paraprotein	0.408 (0.229, 0.588)	0.390 (0.164, 0.617)
Beta2	0.071 (-0.135, 0.277)	0.192 (-0.026, 0.411)
Albumin	-0.361 (-0.596, -0.127)	-0.403 (-0.718, -0.088)
t(4;14): Yes	1.477 (0.958, 1.995)	
Gain(1q21): Yes	0.669 (0.264, 1.074)	0.686 (0.119, 1.254)

Table 3.7: Log-hazard ratio estimates from univariable Cox PH models for TTR and OS applied to the same Myeloma dataset used in the final multivariable models with $n = 125$ to enable comparisons with multivariable models

and albumin 0.207 (-0.088, 0.502). For paraprotein, each SD increase resulted in an increase in the log-HR of 0.310 in the univariable model which was slightly attenuated (0.282) in the model taking into account other covariates. Contrary to the effect in the much reduced dataset, higher levels of beta2 were significantly associated with a higher risk of relapsing in both univariable and multivariable models. Finally, for TTR the negative effect of albumin persisted in both univariable and multivariable models.

In the OS models, log-RD was not significantly associated with the time to death in the univariable models and even after adjusting for the other covariates. Higher levels of paraprotein were predictive of poorer OS in the univariable model. However, this effect was not significant after adjusting for the other covariates. An increase for each SD of beta2 was predictive of a 0.219 and 0.222 increase in the log-HR of death in the univariable and multivariable models respectively. Finally, albumin was again negatively associated with OS in both models as in all previous models. Its effect was however not significant in the multivariable model.

While the analysis thus far has shown that log-RD, paraprotein, beta2, t(4:14) and gain(1q21) are all important in predicting TTR while log-RD, paraprotein, beta2, and

TTR models	Log-HR (95% CI)	Adj. Log-HR (95% CI)
Log-RD	0.352 (0.066, 0.638)	0.207 (-0.088, 0.502)
Paraprotein	0.310 (0.166, 0.454)	0.282 (0.130, 0.434)
Beta2	0.178 (0.044, 0.312)	0.168 (0.024, 0.311)
Albumin	-0.299 (-0.463, -0.135)	-0.248 (-0.421, -0.075)

OS models	Log-HR (95% CI)	Adj. Log-HR (95% CI)
Log-RD	0.288 (-0.114, 0.689)	0.207 (-0.222, 0.636)
Paraprotein	0.201 (0.007, 0.396)	0.175 (-0.034, 0.384)
Beta2	0.219 (0.059, 0.379)	0.222 (0.053, 0.390)
Albumin	-0.248 (-0.472, -0.023)	-0.187 (-0.422, 0.048)

Table 3.8: Log-HR and adjusted log-HR estimates from univariable and multivariable Cox PH models for TTR and OS applied to the Myeloma dataset ignoring t(4:14) and gain(1q21), $n = 243$.

gain(1q21) are important predictors of OS, the missingness in some of the covariates posed challenges, illustrating that focusing only on the CC analysis would result in spurious conclusions as the sensitivity analysis demonstrated. Taking advantage of the various methods that exist for handling missing data, we used multiple imputation (MI) to again model the TTR and OS in Section 3.4.

3.3 Application of standard survival methods to CLL data

To investigate the predictors of TTR and OS in CLL, the methods from Chapter 2 were applied to a second dataset containing records from 415 patients recruited into two clinical trials comparing treatments in CLL. The first trial was Attenuated dose Rituximab with ChemoTherapy In CLL (ARCTIC) - a randomised, phase IIB trial which recruited 200 previously untreated patients with CLL to compare two multiple-drug treatment regimens [107]. The second trial was the ADdition of Mitoxantrone to Improve REsponse to FCR chemotherapy in patients with CLL (ADMIRE), a phase II, multicentre, randomised, controlled, open, parallel group trial of 215 patients [108]. Those without a recorded RD percentage measurement or date when this measure was taken (87) and 3 others with invalid times to event (RD percentage taken before TTR or death) were excluded from the final dataset, leaving 325 patients.

3.3.1 Predictors of TTE outcomes in CLL

Similar to the Myeloma dataset, we summarise the covariates used in modelling the TTR and OS in CLL, Table 3.9. Unlike in the Myeloma trial, there was complete data for most of the variables in CLL except VH mutation which had data from 39 individuals missing, and p-deletion which was not recorded in 24 patients. As in the Myeloma dataset, we start by performing a CC analysis followed by the use of methods for missing data in Section 3.4.

3.3.2 Descriptive summary of the categorical variables in CLL

Summaries based on categorical variables are shown in Table 3.10. Once again, due to differences in the follow-up times, we only comment on the overall numbers in each categorical variable. There were more males, 72%, recruited to the CLL trials than females. In the CLL trials, the Binet staging system was used. It has 3 stages given the

Covariate	Covariate type	Available (%)	Missing (%)
Log-RD	Continuous	325 (100)	0 (0)
Age	Continuous	325 (100)	0 (0)
Sex	Binary	325 (100)	0 (0)
Stage	Categorical	325 (100)	0 (0)
Treatment	Categorical	325 (100)	0 (0)
Trial	Categorical	325 (100)	0 (0)
VH mutation	Binary	286 (88.0)	39 (12.0)
P-deletion	Binary	301 (92.6)	24 (7.4)

Table 3.9: A summary of the covariates used in the analysis of the CLL dataset showing proportion missing for each covariate. The total number of patients in the analysis was 325

letters A, B or C which are dependent on the numbers of red blood cells and platelets in the blood as well as number of areas in the body with enlarged lymphatic tissue [109]. Around 67% of the patients were in stage A or B of the disease and more than 60% of these did not have a mutation of the immunoglobulin variable heavy chain (VH). VH mutation is one of the known risk factors of adverse outcomes in CLL. On the other hand, almost all the patients had a deletion at chromosome 17p-deletion or just p-deletion, which is another important predictor of survival outcomes in CLL [110]. The randomisation to treatment was spread across 4 groups. Finally, a similar number of patients was assigned to both trials.

	TTR		OS		
Variable (n)	Censored	Relapsed	Censored	Died	Total (%)
Sex (325)					
Male	138 (58.8)	97 (41.2)	193 (82.1)	42 (17.9)	235 (72.3)
Female	67 (74.4)	23 (25.6)	79 (87.8)	11 (12.2)	90 (27.7)
Stage (325)					
A or B	137 (62.6)	82 (37.4)	186 (84.9)	33 (15.1)	219 (67.4)
C	68 (64.2)	38 (35.8)	86 (81.1)	20 (18.9)	106 (32.6)
VH mutation (286)					
No	97 (54.5)	81 (45.5)	146 (82.0)	32 (18.0)	178 (62.2)
Yes	80 (74.1)	28 (25.9)	93 (86.1)	15 (13.9)	108 (37.8)
P-deletion (301)					
No	6 (50.0)	6 (50.0)	8 (66.7)	4 (33.3)	12 (4.0)
Yes	181 (62.6)	108 (37.4)	243 (84.1)	46 (15.9)	289 (96.0)
Treatment (325)					
Control	103 (65.2)	55 (34.8)	132 (83.5)	26 (16.5)	158 (48.6)
ADMIRE Exp	49 (56.3)	38 (43.7)	72 (82.8)	15 (17.2)	87 (26.8)
ARCTIC Exp 1	37 (61.7)	23 (38.3)	49 (81.7)	11 (18.3)	60 (18.5)
ARCTIC Exp 2	16 (80.0)	4 (20.0)	19 (95.0)	1 (5.0)	20 (6.2)
Trial (325)					
ADMIRE	100 (58.8)	70 (41.2)	138 (81.2)	32 (18.8)	170 (52.3)
ARCTIC	105 (67.7)	50 (32.3)	134 (86.5)	21 (13.5)	155 (47.7)

Table 3.10: Summary of categorical variables in the CLL dataset by TTR and OS outcomes. For VH mutation*, $n = 286$, while $n = 301$ for p-deletion**

.

3.3.3 Summary of the continuous variables by TTR and OS

Boxplots of the two continuous variables in the CLL dataset are shown in Figure 3.7. The log-RD measurements were positively skewed with a median [IQR] of -2.40 [-2.40, 1.57] The median age for all patients at randomisation was 63.3 years, IQR [57.2, 68.1 years].

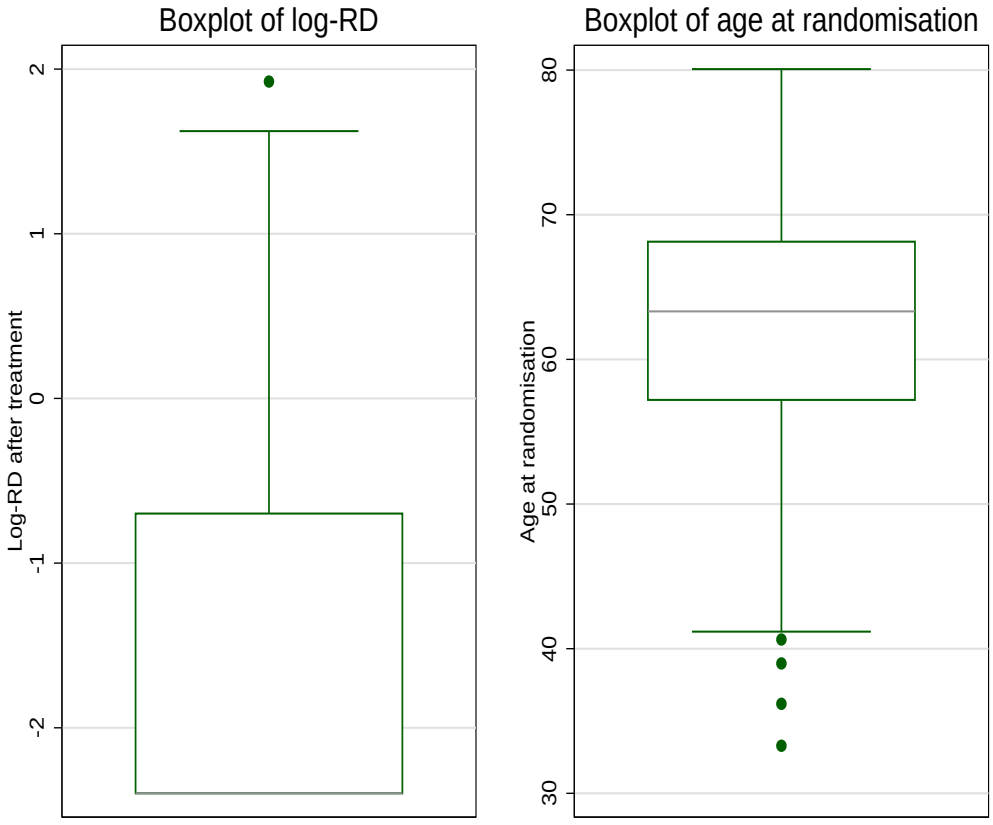


Figure 3.7: Box plots for log-RD and age at randomisation in the CLL dataset

3.3.4 Overall TTR and OS patterns in CLL

The median TTR for all patients was 2 years inter-quartile range (IQR) [1.7, 2.3 years]. On top of this, relapses appeared to occur over the who range of follow-up. The median OS after treatment was 6.6 years. Over the 7 years follow-up, around 40% were still alive, Figure 3.2. The OS curve was less steep than the TTR curve as CLL is disease patients can live with for several years even if they are not cured.

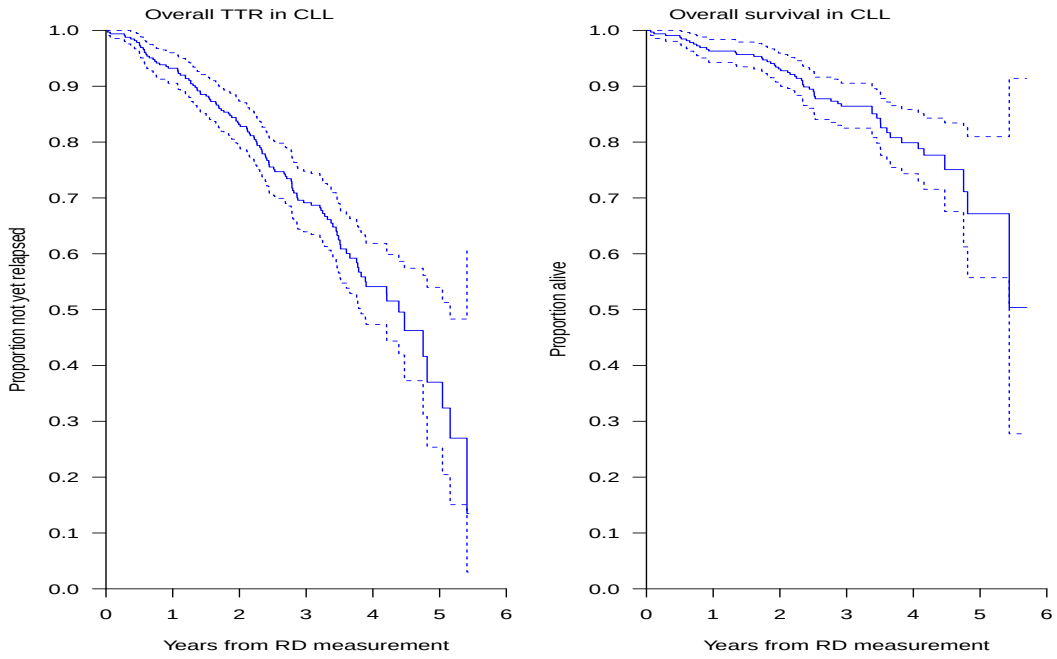


Figure 3.8: K-M estimates of the overall TTR (left) and OS (right) in the CLL dataset. Dotted lines represent 95% CI limits.

3.3.5 Comparing TTR patterns by categorical variables in CLL

K-M curves were used to compare the TTR patterns using the categorical variables, Figure 3.9. The TTR curves were continually decreasing over the follow-up time. In contrast to the survival patterns observed in the Myloma IX patients, the TTR curves did not have discernible plateaus, suggesting that identification of a proportion of patients who are cured, by the definition of no relapse, may be difficult. There were no obvious differences in the relapse patterns by sex, disease stage, trial (ARCTIC versus ADMIRE) and treatment. The few without 17p-deletion had lower relapse rates than the majority who had 17p-deletion. Those with VH mutation had better outcomes than those without a VH mutation.

3.3.6 Comparing OS patterns by categorical variables in CLL

OS patterns were also plotted by the categorical variables. Overall, more than half the patients were still alive 5 years after the RD percentage measurement, Figure 3.10. There were no obvious differences in OS by trial, treatment or stage and there were similar OS patterns by sex and VH mutation risk. Those with p-deletion had better OS than the few without p-deletion.

3.3.7 Modelling independent factors associated with TTR in CLL

Univariable Cox PH, Exponential and Weibull regression models were fitted to the CLL data to ascertain which variables were independently associated with TTR. Estimates of the log-HRs as well as the 95% CIs from these three models relating each covariate and the TTR are shown in Table 3.11.

In terms of factors associated with TTR in CLL, there were no significant differences in the relapse probabilities by disease stage. Those with more advanced disease at recruitment, stage C, had a lower risk of relapsing after RD measurement log-HR (95%

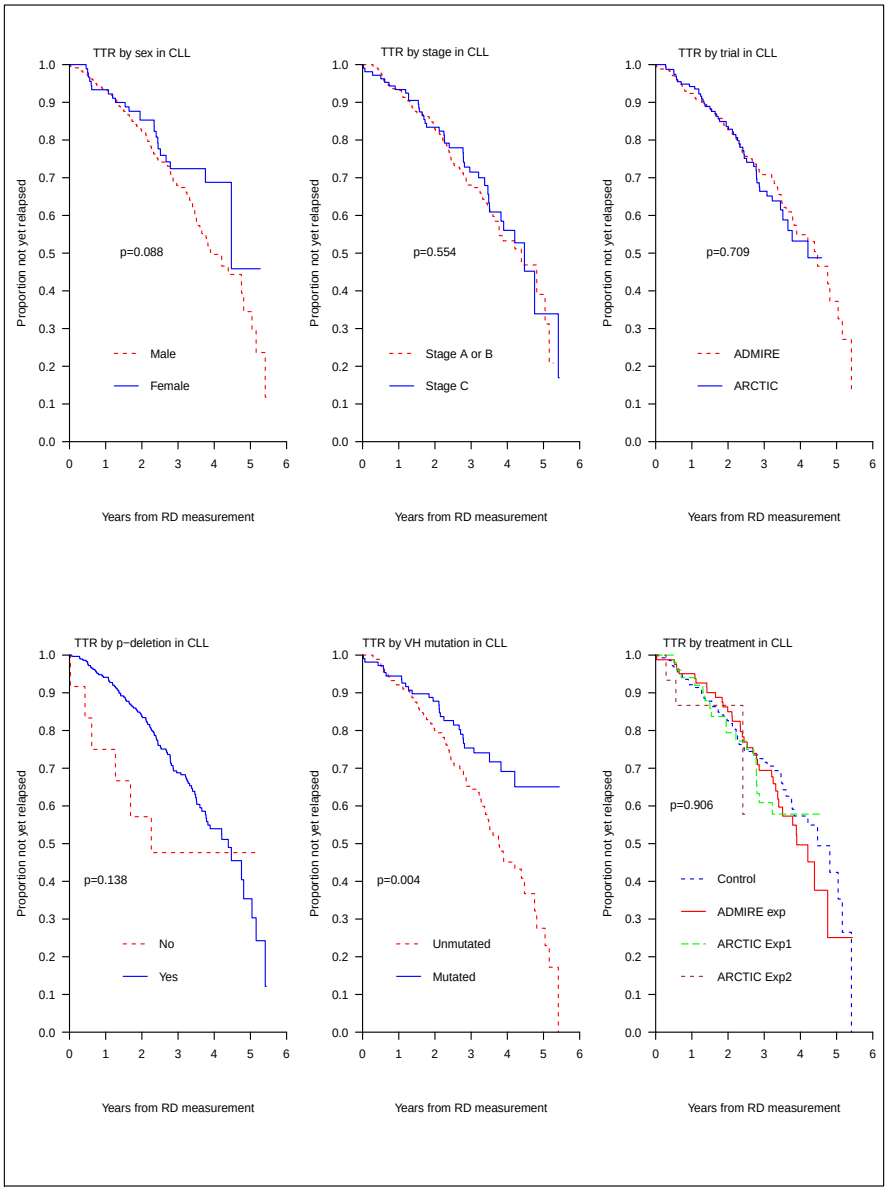


Figure 3.9: K-M plots for TTR by categorical variables in the CLL dataset

CI) -0.117 (-0.506, 0.272) when compared to those with stage A or B disease, confirming the TTR patterns depicted by the K-M survivor curves. Likewise, estimates of the log-HR from the Exponential and Weibull models while slightly lower, showed no effect of stage on TTR. In all models, there was no effect of age at randomisation on

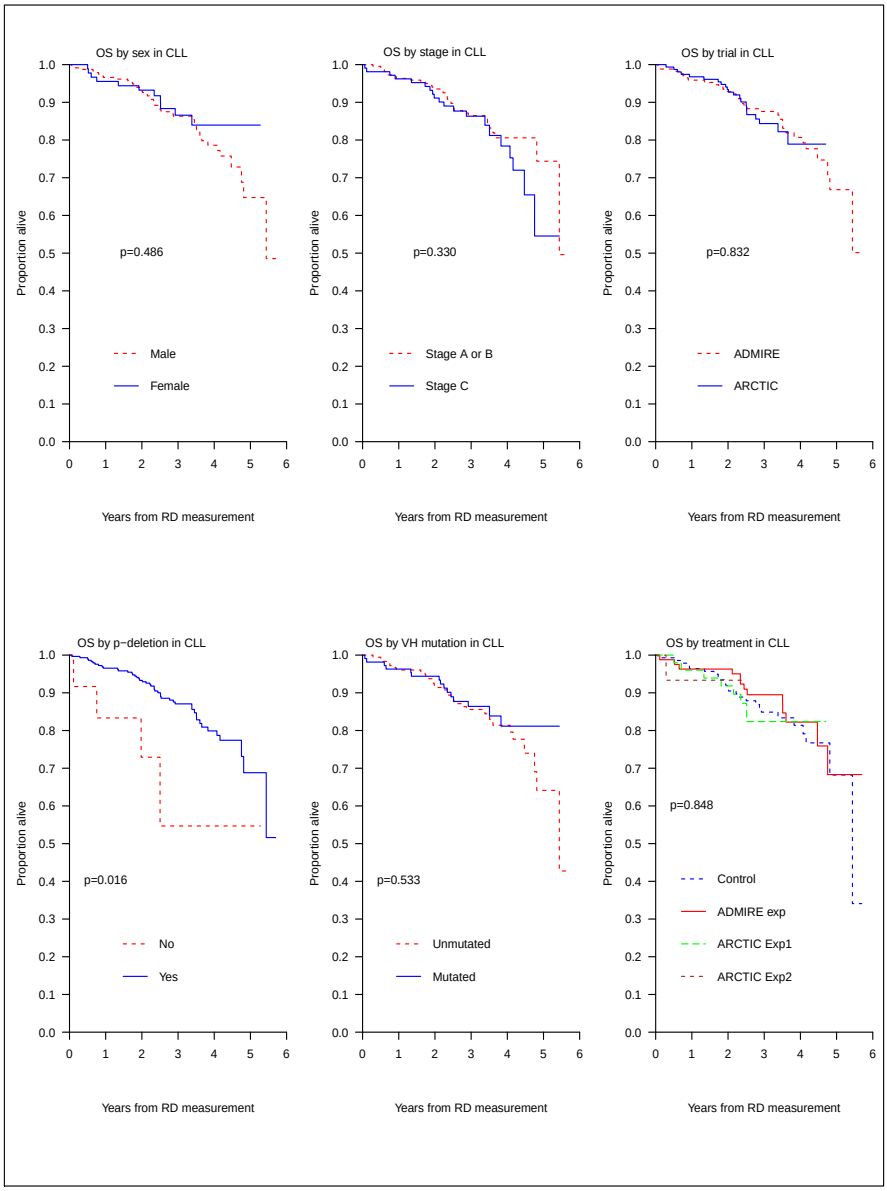


Figure 3.10: K-M plots for OS by categorical variables in the CLL dataset

the TTR. The log-RD was an important predictor of TTR for these CLL patients. Each log-increase in the RD percentage after treatment was associated in a 0.819 (0.670, 0.967) increase in the log-HR. Again, there were no differences in TTRs among males and females while those without VH mutation had significantly longer TTR after treat-

ment, -0.626 (-1.056, -0.195). Finally, p-deletion was not associated with TTR for these patients. It is worth noting that the majority had p-deletion while only (12/301) were in the other group without p-deletion.

The shape parameter in the Weibull models was greater than 1 which implies that the relapse rate increased with time. However, as with the Myeloma data, we focus on estimates from the more flexible Cox PH models.

Test for the PH assumption in univariable models for TTR in CLL

Once again, plots of the scaled Schoenfeld residuals against time for all the covariates in the CLL dataset were used to assess the validity of the PH assumption after fitting the univariable Cox PH models for TTR, Figure 3.11. From the plots, the PH assumption generally held for most of the covariates used in the univariable Cox PH models for TTR. Further, the p-values from the global test for the PH assumption and for all covariates in univariable models were > 0.05. These plots were made based on the available data ignoring missing values.

	Cox PH	**Exponential**	**Weibull**
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Intercept		-2.012 (-2.228, -1.796)	-2.674 (-3.042, -2.305)
Stage C	-0.117 (-0.506, 0.272)	-0.058 (-0.443, 0.326)	-0.080 (-0.465, 0.305)
Shape (γ)			1.575 (1.342, 1.847)
Intercept		-2.291 (-3.694, -0.887)	-2.932 (-4.371, -1.492)
Age	0.001 (-0.022, 0.023)	0.004 (-0.018, 0.026)	0.004 (-0.019, 0.026)
Shape (γ)			1.573 (1.340, 1.845)
Intercept		-1.165 (-1.374, -0.956)	-1.933 (-2.283, -1.582)
Log-RD	0.819 (0.670, 0.967)	0.675 (0.537, 0.813)	0.788 (0.642, 0.933)
Shape (γ)			1.809 (1.554, 2.105)
Intercept		-1.929 (-2.128, -1.730)	-2.595 (-2.957, -2.234)
Sex: Female	-0.394 (-0.850, 0.062)	-0.444 (-0.899, 0.010)	-0.423 (-0.878, 0.031)
Shape (γ)			1.568 (1.336, 1.840)
Intercept	-	-1.809 (-2.027, -1.591)	-2.485 (-2.868, -2.101)
VH mutation*: Yes	-0.626 (-1.056, -0.195)	-0.615 (-1.045, -0.186)	-0.639 (-1.069, -0.209)
Shape (γ)			1.583 (1.339, 1.870)
Intercept		-1.405 (-2.205, -0.605)	-2.026 (-2.877, -1.175)
P-deletion**: Yes	-0.620 (-1.452, 0.211)	-0.617 (-1.439, 0.205)	-0.683 (-1.506, 0.139)
Shape (γ)			1.591 (1.352, 1.872)

Table 3.11: Estimates of log-HRs from univariable Cox PH, Exponential and Weibull models for TTR in the CLL dataset. In the Exponential and Weibull models, estimates of the intercept in the model for the log-hazard are presented on top of the log-HR, while the shape parameter (γ) is also estimated in the Weibull models. There were 286 patients with VH mutation* records and 301 with p-deletion**. For the other covariates, models were fitted with $n = 325$ as they had no missing data.

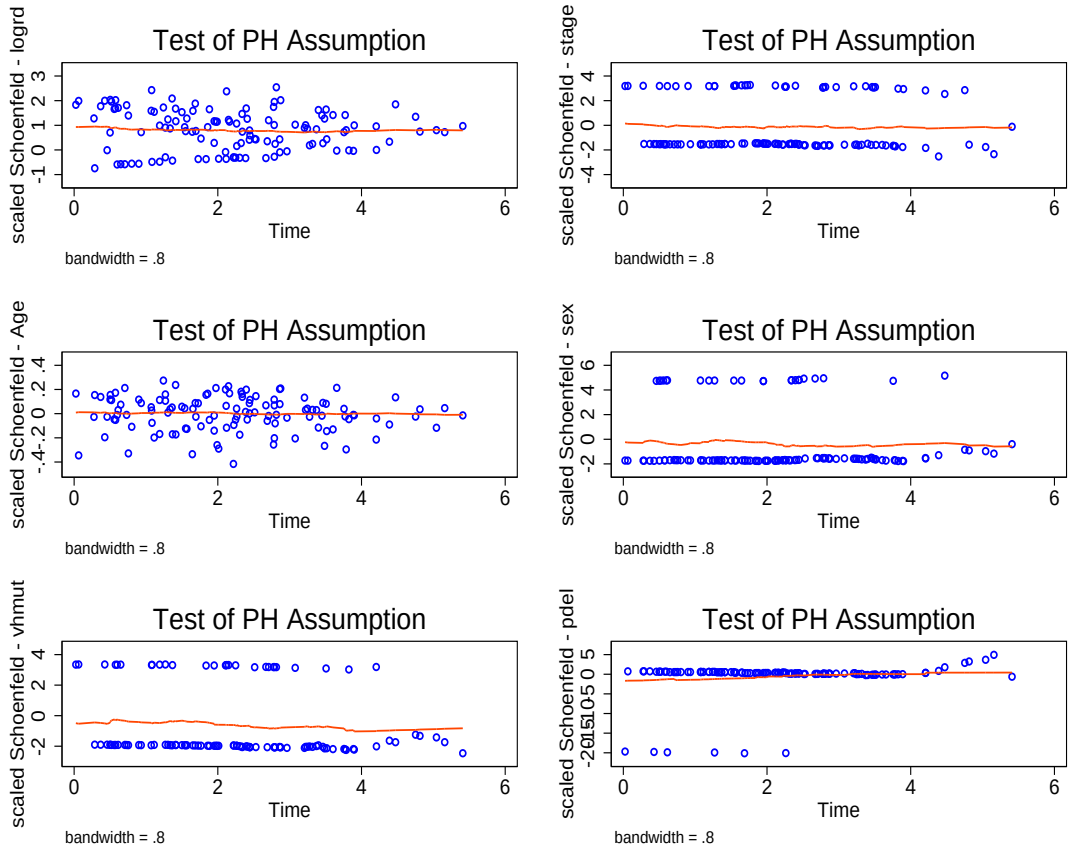


Figure 3.11: Scaled Schoenfeld residuals plotted against time from univariable Cox PH models for TTR in the CLL dataset ignoring missing data

3.3.8 Modelling independent factors associated with OS in CLL

Next we looked at independent factors associated with OS in CLL by fitting univariable Cox PH, Exponential and Weibull models. Estimates of the log-HRs, their 95% CIs and the shape parameter (γ) from the Weibull model are shown in Table 3.12. Once again, a CC analysis was done for VH mutation (286 patients) and p-deletion (301 patients). Patients with higher log-RD were at a significantly higher risk of death than those with low RD burden after treatment. Each log-increase in the RD percentage after treatment was associated with a significant increase of 0.452 (0.247, 0.657) in the

log-hazard of dying. Patients with stage C disease had a higher risk of death than stage A or B patients although the effect was not significant, log-HR (95% CI), 0.275 (-0.281, 0.831). The age at randomisation was borderline significant in the Cox PH model with a increase in the log-HR of 0.035 for each year. A similar effect of the age on OS was seen in the Exponential and Weibull models. Sex and VH mutation risk were not significantly associated with OS. Finally, patients with p-deletion had significantly longer OS. It is worth noting that there were only 12 patients without p-deletion so that this effect could well have been superficially big.

Test for the PH assumption in univariable models for OS in CLL

Again the scaled Schoenfeld residuals were used to test the PH assumption in the univariable Cox PH models for OS, Figure 3.12. The proportionality assumption was again formally tested for each covariate. The residuals plots for all covariates from the univariable Cox PH models for OS showed no significant departures from the PH assumption. Moreover, all p-values from the formal test for the PH assumption for each covariate were > 0.05.

	Cox PH	**Exponential**	**Weibull**
Parameter	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Intercept		-3.040 (-3.382, -2.699)	-3.607 (-4.173, -3.042)
Stage C	0.275 (-0.281, 0.831)	0.257 (-0.298, 0.813)	0.262 (-0.294, 0.817)
Shape (γ)			1.459 (1.141, 1.867)
Intercept		-5.064 (-7.319, -2.808)	-5.717 (-8.029, -3.405)
Age	0.035 (0.000, 0.070)	0.033 (-0.001, 0.068)	0.035 (0.000, 0.069)
Shape (γ)			1.468 (1.148, 1.878)
Intercept		-2.392 (-2.718, -2.065)	-2.966 (-3.516, -2.417)
Log-RD	0.452 (0.247, 0.657)	0.445 (0.240, 0.650)	0.459 (0.253, 0.665)
Shape (γ)			1.483 (1.162, 1.892)
Intercept		-2.878 (-3.180, -2.575)	-3.443 (-3.989, -2.896)
Sex: Female	-0.237 (-0.906, 0.432)	-0.312 (-0.975, 0.352)	-0.281 (-0.945, 0.384)
Shape (γ)			1.453 (1.135, 1.859)
Intercept		-2.877 (-3.224, -2.531)	-3.347 (-3.923, -2.771)
VH mutation*: Yes	-0.195 (-0.811, 0.420)	-0.223 (-0.836, 0.391)	-0.212 (-0.825, 0.402)
Shape (γ)			1.375 (1.057, 1.790)
Intercept		-1.912 (-2.892, -0.932)	-2.353 (-3.408, -1.299)
P-deletion**: Yes	-1.198 (-2.228, -0.168)	-1.069 (-2.091, -0.047)	-1.153 (-2.176, -0.129)
Shape (γ)			1.426 (1.108, 1.835)

Table 3.12: Estimates of log-HRs from univariable Cox PH, Exponential and Weibull models for OS in the CLL dataset. In the Exponential and Weibull models, estimates of the intercept in the model for the log-hazard are presented on top of the log-HR, while the shape parameter (γ) is also estimated in the Weibull models. There were 286 patients with VH mutation* records and 301 with p-deletion**. For the other covariates, models were fitted with $n = 325$ as they had no missing data.

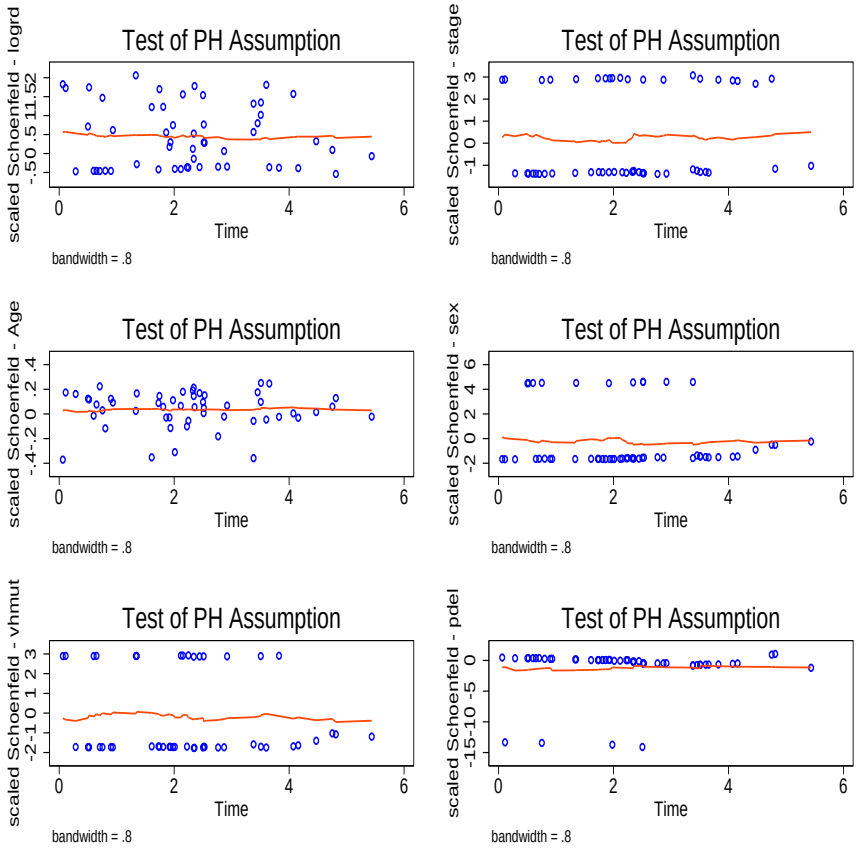


Figure 3.12: Scaled Schoenfeld residuals plotted against time from univariable Cox PH models for OS in the CLL dataset ignoring missing data

3.3.9 Multivariable models for TTR and OS in CLL

Multivariable Cox PH, Exponential and Weibull models were fitted to examine the effect of the log-RD while taking into account other covariates. The same model selection approach used in Myeloma was applied in Stata, starting with the Cox PH model and then fitting the same covariates in the final model to Exponential and Weibull models. The results are shown in Table 3.13.

The final model for TTR was based on data from 286 individuals who had log-RD, disease stage and VH mutation recorded. Those with higher log-RD had shorter TTR

when stage and VH mutation were taken into account. Each log-increase in the RD percentage at baseline was associated with a significant 0.849 (0.687, 1.011) increase in the log-HR. Presence of a VH mutation was associated with a lower risk of relapsing as in the univariable models although this was not significant. The effect of stage was likewise not significant when log-RD and VH mutation were taken into account. The estimates from the Cox PH and Weibull models were generally similar while those from the Exponential model were in some cases slightly different.

The final model for OS included log-RD, stage and p-deletion and was based on data from 301 patients in a CC analysis. Again a higher disease burden was predictive of poor OS with similar estimates across all models. Each log-increase in the RD percentage after treatment was significantly associated with a 0.406 (0.193, 0.619) increase in the log-HR in the multivariable Cox PH model. Those with stage C disease were at a higher risk of dying than those in stages A or B although this was not significant. Having p-deletion was associated with improved OS chances, although this effect was not significant.

The estimates across the various multivariable models for both TTR and OS were generally similar while the Weibull model had a shape parameter significantly greater than 1, suggesting an increase in the hazard rate with time. The multivariable Weibull model also had the smallest AIC for both the TTR and OS meaning it performed better than the others. However, an inspection of the Schoenfeld residuals showed that the PH assumption held in both multivariable models for TTR and OS applied to the CLL dataset as shown in Figure F.1 and Figure F.2 in Appendix F. It would thus suffice to model the CLL data using the Cox PH model.

As a way of enabling comparisons, we fit univariable Cox PH models for TTR with the 286 patients who were included in the final adjusted model and the 301 patients in the final OS model, focusing only on the covariates in these final models, Table 3.14. For TTR, the effect of log-RD remained significant in both univariable and multivariable

	Cox PH	**Exponential**	**Weibull**
Parameter	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
TTR model			
Intercept	-	-0.947 (-1.223, -0.671)	-1.689 (-2.083, -1.295)
Adjusted log-HRs			
Log-RD	0.849 (0.687, 1.011)	0.669 (0.522, 0.815)	0.811 (0.654, 0.969)
Stage C	-0.282 (-0.695, 0.130)	-0.114 (-0.518, 0.290)	-0.212 (-0.618, 0.193)
VH mutation: Yes	-0.651 (-1.090, -0.212)	-0.544 (-0.976, -0.112)	-0.668 (-1.103, -0.232)
Shape (γ)	-		1.873 (1.599, 2.195)
AIC	984.1	458.1	413.8
OS model			
Intercept	-	-1.729 (-2.733, -0.725)	-2.128 (-3.186, -1.071)
Adjusted log-HRs			
Log-RD	0.406 (0.193, 0.619)	0.399 (0.185, 0.612)	0.410 (0.196, 0.625)
Stage C	0.257 (-0.317, 0.832)	0.228 (-0.344, 0.800)	0.243 (-0.330, 0.816)
P-deletion: Yes	-1.023 (-2.065, 0.019)	-0.815 (-1.847, 0.216)	-0.968 (-2.003, 0.067)
Shape (γ)	-		1.454 (1.132, 1.868)
AIC	500.9	330.3	324.9

Table 3.13: Estimates of adjusted log-HRs from Cox PH, Exponential and Weibull models for TTR and OS in the CLL dataset. The final TTR model was fitted to data from 286 individuals while the OS model had data from 301 patients

	Cox PH for TTR	**Cox PH for OS**
	Log-HR (95% CI)	Log-HR (95% CI)
Log-RD	0.805 (0.652, 0.959)	0.428 (0.217, 0.639)
Stage C	-0.123 (-0.530, 0.284)	0.321 (-0.250, 0.893)
VH mutation: Yes	-0.626 (-1.056, -0.195)	
P-deletion: Yes		-1.198 (-2.228, -0.168)

Table 3.14: Univariable Cox PH models for TTR and OS in the CLL dataset fitted to data from 286 individuals in the multivariable TTR model and the 301 patients in the final OS model respectively

models with a bigger effect when stage and VH mutation are taken into account log-HR 0.805 in the univariable model, versus 0.849 in the adjusted analysis. Stage and VH mutation were both not associated with the TTR in both models. In the OS models, log-RD was associated with OS in both the unadjusted and adjusted models. However, the effect in the multivariable model was slightly smaller. The effect of stage remained non-significant in both models, just like that of p-deletion. The CC analysis ignoring the missingness might again lead to wrong conclusions. We thus used methods for handling missing data to deal with this problem in Section 3.4.

3.4 Handling missing data

In the models fitted in this analysis so far, we only considered individuals with data on all the variables included in the models in a CC analysis. In situations where data are *missing completely at random* (MCAR) in the terminology of [111], CC analysis produces unbiased results [112]. However, if data are missing at random (MAR), conditional on data that we do observe (say age and RD) we can fit models with multiply imputed data in order to obtain less biased estimates. Imputed data are based on imputation models, usually regression models based on observed predictors of the missing data. However, for more complicated missing data mechanisms, that may be missing not at random (MNAR) we cannot adjust for missing data using the complete cases and sensitivity analysis is required. To investigate whether estimates from models based on the CC analysis are similar to those based on the overall data likelihood as given in Equation 2.23, MI techniques that have been widely discussed in the literature can be used, including those specifically developed for handling missing data in Cox PH models [113]. We restrict attention in this work to that case where the data are missing at random (MAR) and following the work of others use the three prescribed stages of MI which involve:

1. Generating m imputed datasets where the missing data are replaced by the posterior predictive distribution of the missing values given the observed data.

2. Fitting the analysis models using the data from the resulting 'complete' datasets, and

3. Combining the estimates from the m datasets using Rubin's rules [114].

As data maybe be missing across several variables in an analysis, an approach called multiple imputation using chained equations (MICE) has become the method of

choice for imputing the missing values in the m datasets. MICE is based on a set of imputation models, one for each covariate with missing data [115]. Briefly, MICE performs MI by regressing the first variable with missing values on all other variables that may include the outcome variable, but only on individuals with data on the first variable. The missing variables in this particular variable are then replaced by simulated draws from its predictive distribution. Both imputed and observed values for the first variable, alongside other observed data are used to build an imputation model (regression) for the next variable, to provide imputations for missing values of variable the second variable. This is repeated for all variables with missing values to complete one cycle. Several cycles may be repeated (analogous to the MCMC approach) to produce a single complete imputed data set. This procedure is repeated to produce the m datasets.

To impute missing values in the Myeloma and CLL datasets, we used MICE in Stata with $m = 10$ (10 datasets with missing covariates values imputed). In this case we chose $m = 10$ as it has been shown that 5 to 10 imputed datasets would normally suffice, while bearing in mind that m should ideally depend on the percentage of missing data [116]. Further, we used linear regression model as the imputation model for continuous covariates with missing data, and logistic regression for binary covariates with missing data. We included auxiliary variables that were not part of the analysis model such as age and sex to help improve prediction of the missing values [117], as well as the observed log-transformed TTR and/or OS including the censoring indicator. In this way, we ensured that we utilise all the available data, including the outcome and censoring indicators to help inform the imputation, [118]. To deal with the potential problem with perfect prediction in the logistic regression imputation models for binary variables, we used augmented logistic regression which adds extra observations into the dataset to deal with the problem.

3.4.1 Missing data in the Myeloma dataset

As summarised in Section 3.2.1, the Myeloma data had several variables with missing data. We first looked at the missing data patterns to determine if there were systematic ways in which the data was missing, Table 3.15. In this dataset, there was complete information for 29% (119/427) of the patients. Further, there were a total of 19 different patterns of missingness across the 7 covariates with missing information. From this table, hyperdiploidy, t(4;14), t(11;14) and gain(1q21) were missing together 42% of the time. Otherwise, it was not evident that the missing data followed a particular pattern across all covariates. A formal check was done to determine if the fact that a variable was missing for each covariate were related to the other covariates with available data (age, sex, TTR, censoring and log-RD) using a logistic model with (1 if value missing and 0 if not missing). This analysis showed that the probability of a missing observation was not related to the other fully observed covariates (results not shown). We thus proceed to perform MI using MICE in Stata using linear regression for missing continuous covariates and logistic regression for binary covariates respectively as the imputation models.

3.4.2 Baseline characteristics of participants with complete and incomplete data in the Myeloma dataset

After the imputation which resulted in 10 datasets, we compared the means, for continuous variables, and proportions for categorical variables, between the available data used in the CC analysis, and the summaries across 10 imputed datasets for the missing data as shown in Table 3.16. There was a higher proportion of patients with hyperdiploidy in those with complete data than in those with missing data (53.3% vs 47.5%). The translocations t(4:14) and t(11:14) were more prevalent among the incomplete cases after MI compared to the completely observed cases. Gain(1q21) was

Number of patients	Percent	Paraprotein	Beta2	Albumin	Hyperdiploidy	t4;14	t11;14	Gain(1q21)
123	29	+	+	+	+	+	+	+
98	23	+	+	+	-	-	-	-
54	13	+	-	+	+	+	+	+
34	8	-	+	+	-	-	-	-
33	8	+	-	+	-	-	-	-
28	7	-	+	+	+	+	+	+
15	4	+	+	+	+	+	+	-
14	3	-	-	+	+	+	+	+
11	3	-	-	+	-	-	-	-
4	< 1	-	+	+	+	+	+	-
2	< 1	+	-	+	+	+	+	-
2	< 1	+	+	+	-	+	+	-
2	< 1	+	+	+	-	+	+	+
1	< 1	-	-	+	+	+	-	-
1	< 1	-	+	+	-	+	+	-
1	< 1	+	-	-	+	+	+	+
1	< 1	+	-	+	+	+	-	-
1	< 1	+	+	+	+	-	-	-
1	< 1	+	+	+	+	-	+	-
1	< 1	+	+	+	+	+	-	-
427	100							

Table 3.15: Missingness patterns in the Myeloma dataset. The (+) represents observed data while the (-) represents missing data.

Variable	n_{cc}	Complete cases	n_{ic}	Incomplete cases
Hyeperdiploidy (Yes), fraction (%)	246	131 (53.3)	181	86 (47.5)
t4;14 (Yes), fraction (%)	249	35 (14.1)	178	28 (15.7)
t11;14 (Yes), fraction (%)	247	40 (16.2)	180	41 (22.8)
Gain(1q21) (Yes), fraction (%)	222	80 (36.0)	205	70 (34.1)
Paraprotein, mean (SD)	334	1.3e-8 (1.00)	93	-0.03 (1.07)
Beta2, mean (SD)	310	-1.7e-9 (1.00)	117	0.04 (1.03)
Albumin, mean (SD)	426	-4.9e-10 (1.00)	1	0.11 (0.00)

Table 3.16: Comparison of incomplete and completed binary and continuous variables in the Myeloma dataset. Note: SD for albumin not calculated as only one value was missing. n_{cc} represents number of complete cases while n_{ic} is the number of incomplete cases. For the incomplete cases, the means and proportions are averaged over the 10 datasets.

more prevalent among the complete cases compared to the incomplete cases (36% vs 34%). The means and SDs for the standardised paraprotein, beta2 and albumin were higher among those with missing data compared to those who had these measures although the distributions were still approximately standard Normal.

We plotted histograms of the original values of the z-standardised paraprotein and beta2 and overlaid the histograms with those from the first imputed dataset (grey). Further, we plotted boxplots of the original z-standardised paraprotein and beta2 together with those for these measures from each of the 10 imputed datasets, Figure 3.13. The distributions of the imputed values spanned the whole range of values compared to the observed cases which were positively skewed for both paraprotein and beta2 as seen from the boxplots. The medians were higher in the imputed datasets for both paraprotein and beta2 and imputed values of paraprotein generally fell within the IQR for most of the datasets. For beta2, some imputed values tended to fall at the lower end of distribution. Based on this, we can infer that we do not have MCAR, but can instead reasonably assume MAR.

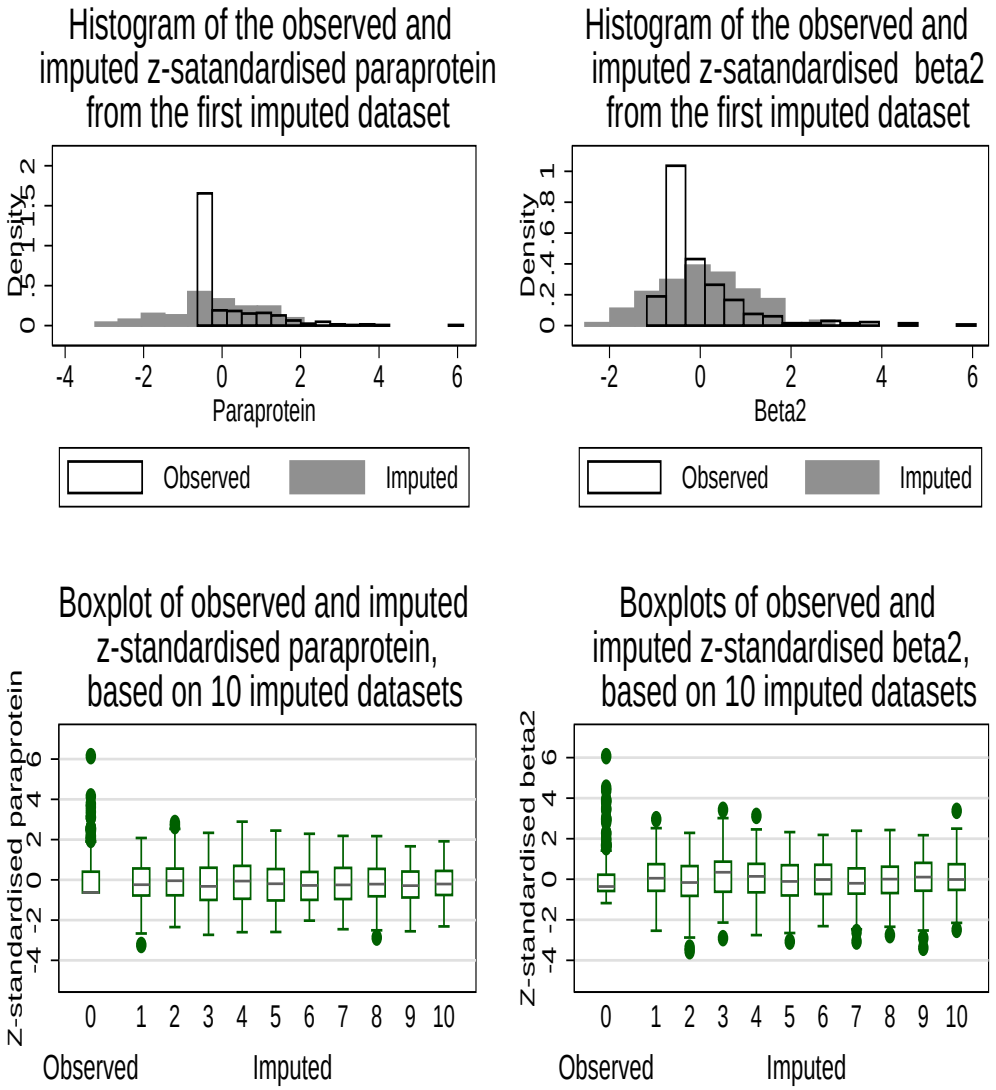


Figure 3.13: Histograms of z-standardised paraprotein and beta2 overlaid by imputed values from the first imputation (grey) and boxplots showing the distributions of the observed and imputed z-standardised paraprotein and beta2 from 10 imputed datasets (bottom)

3.4.3 Univariable Cox PH models for TTR and OS in Myeloma fitted to imputed data

We then fit univariable Cox PH models for both TTR and OS to the imputed datasets and combined the results using standard Stata routines. For comparison, we looked at the

estimates from the univariable models ignoring missing data. We reported estimates of the log-HRs, p-values (p) and the fraction of missing information (FMI) for each model, Table 3.17. The FMI may be used to determine the gains from using MI as it is used to quantify the loss of information due to the missing data while accounting for the information retained by the other variables within the dataset. We would expect the FMI to be less than the proportion of missing data for a given variable as it takes into account the predictive power of the given imputation model. FMI values range between 0 and 1 with values closer to 1 indicative of high variability between imputed datasets such that the observed data in the imputation model does not provide much information about the missing values [119].

The results from the univariable models for TTR showed similar effects to the CC results for paraprotein (positive significant effect), and albumin (negative significant effect) on the TTR with more or less similar standard errors (SEs) and low FMI. In both models, a higher value of beta2 was associated with an increase in the log-HR of relapsing although the estimate with MI was smaller (0.179 v 0.193) compared to the CC model. The effects of hyperdiploidy and t(11;14) remained non-significant even after MI. Those with t(4;14) were at a significantly higher risk of relapsing than those without it. It is worth noting that the FMI from the imputation was bigger, 0.739 and 0.600 for t(4;14) and t(11;14) respectively, meaning the imputation might not have provided enough information about the missingness in these variables. This was also evident from the larger SEs in the imputed models for these variables. Finally, gain(1q21) remained a significant predictor of quicker relapse for those who had this marker. We checked the validity of the PH assumption for all models and present the Schoenfeld residuals from the model fitted to the first imputed dataset, Figure 3.14. For this imputation, the PH assumption held for paraprotein and t(11;14), otherwise the p-value from the global test was significant for beta2, hyperdiploidy, t(4;14) and gain(1q21). This could have well been due to chance.

	CC models		Models with MI		
	Log-HR (SE)	p	Log-HR (SE)	p	FMI
TTR models					
Paraprotein	0.203 (0.055)	< 0.001	0.200 (0.054)	< 0.001	0.175
Beta2	0.193 (0.060)	0.001	0.179 (0.068)	0.012	0.432
Albumin	-0.138 (0.053)	0.009	-0.137 (0.053)	0.009	0.001
Hyperdiploidy: Yes	-0.252 (0.145)	0.082	-0.169 (0.147)	0.258	0.448
t(4;14): Yes	1.284 (0.196)	< 0.001	1.191 (0.271)	< 0.001	0.739
t(11;14): Yes	-0.196 (0.199)	0.326	-0.377 (0.231)	0.114	0.600
Gain(1q21): Yes	0.590 (0.156)	< 0.001	0.540 (0.148)	0.001	0.427
OS models					
Paraprotein	0.093 (0.081)	0.252	0.114 (0.074)	0.126	0.127
Beta2	0.278 (0.068)	< 0.001	0.280 (0.078)	0.001	0.405
Albumin	-0.156 (0.072)	0.031	-0.158 (0.072)	0.030	0.001
Hyperdiploidy: Yes	-0.673 (0.210)	0.001	-0.654 (0.190)	0.001	0.313
t(4;14): Yes	0.889 (0.239)	< 0.001	0.894 (0.237)	< 0.001	0.466
t(11;14): Yes	-0.310 (0.298)	0.298	-0.429 (0.326)	0.198	0.570
Gain(1q21): Yes	0.707 (0.213)	0.001	0.729 (0.203)	0.001	0.448

Table 3.17: Log-HR estimates and SEs, in brackets, from CC univariable Cox PH models fitted to covariates with missing data in the Myeloma dataset and equivalent results obtained via MI in Stata including the fraction of missing information (FMI) for each covariate.

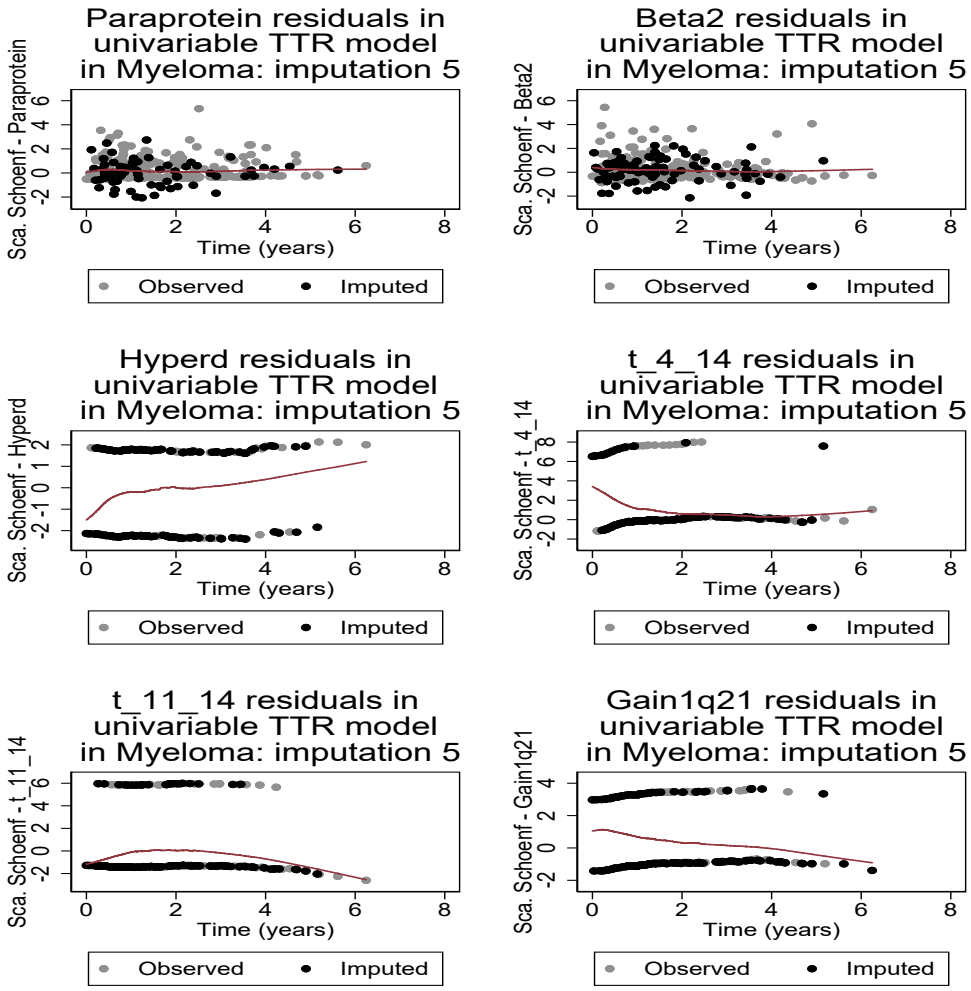


Figure 3.14: Schoenfeld residuals from univariable Cox PH models for TTR with z-standardised paraprotein, beta2, hyperdiploidy (hyperdip), t(4:14), t(11:14) and gan(1q21) as covariates, plotted against time since RD measurement using the first imputed dataset

3.4.4 Multivariable Cox PH models for TTR and OS in Myeloma fitted to imputed data

To obtain the final multivariable model, we performed the same model selection procedure to each of the 10 datasets and selected covariates appearing together in the final

model most of the time. Several formal approaches for model selection within MI have been proposed some of which include forward selection procedures for models with time dependent covariates [120]. For our purposes with 10 imputed datasets, it was feasible to perform model selection for each dataset. We therefore developed a model in each of the 10 datasets and used these to decide on which were the most important covariates with respect to both TTR and OS. Following this, we re-fit the model with the common covariates identified to each of the 10 datasets and then combined the estimates from these models using Rubin's rules. While this approach is practical, it might not work well where the number of imputed datasets is large. Ideally, one would just make inference based on covariates without performing model selection. Results from the final multivariable model for TTR and OS in Myeloma are shown in Table 3.18.

For TTR, the multivariable model included log-RD, paraprotein, beta2, albumin, t(4;14) and gain(1q21). With the MI ensuring that we make inference based on data from all 427 individuals, each log-increase in the log-RD was associated with a 0.107 (-0.125, 0.339) in the log-HR of a relapse when all the other covariates were taken into account. This effect was less than that in the univariable model 0.214 (0.008, 0.421) and also in the final model arrived at in the CC analysis. Higher levels of paraprotein were predictive of higher relapse chances in both univariable and multivariable models with the full data. The effect increased from 0.200 in the univariable model to 0.253 when adjusting for other factors. Its FMI indicated that 23.9% of the sampling variance was attributable to the missing data. As with paraprotein, higher levels of beta2 were significantly associated with TTR. However, its effect was smaller in the multivariable model. The variance attributed to the missing data for beta2 was 36.7% which was high as expected because only 310 patients had data on this marker. Albumin continued to have an inverse relationship with TTR. The effect was however smaller and not significant in the final model based on the full dataset. Finally t(4;14) and gain(1q21) were both significantly associated with shorter TTRs in both univariable and multivariable

models. However, as expected, the FMIs for these variables were very high as they had large proportions of missing values.

The multivariable model for OS in Myeloma included log-RD, paraprotein, beta2, albumin, hyperdiploidy, t(4;14) and gain(1q21). In the multivariable model, log-RD predicted poorer OS although the effect was not significant. The log-HR for log-RD in this final model was less than that in the univariable Cox PH model for OS, 0.087 (-0.203, 0.377) in Table 3.5 of Section 3.2.8, and it was not significant in both models. All other covariates in this final model except albumin and t(4;14) were significantly associated with OS. The effects for paraprotein, beta2, albumin, hyperdiploidy, t(4;14) and gain(1q21) were generally smaller in absolute value when all other covariates were taken into account except for paraprotein.

This analysis enabled us to identify log-RD, paraprotein, beta2, albumin, t(4;14) and gain(1q21) as the important predictors of the TTR. While log-RD is a known predictor of TTR in Myeloma, there were other growth related variables that might play a key role in determining outcomes such as paraprotein whose effect increased when the other covariates were taken into account. While the effect of beta2 was slightly attenuated in the multivariable analysis, it remained another important factor associated with TTR. Related to beta2 was albumin, which together with beta2 are used to determine disease stage. On top of this, presence of the translocation t(4;14) and gain(1q21) were important in predicting TTR in this dataset. This shows that to fully characterise the TTR in Myeloma, we need to consider other mechanisms other than the residual disease alone. These other covariates are markers of tumour growth and we will therefore examine them further to ascertain if they act together in predicting the TTR. For OS, hyperdiploidy was an additional important predictor on top of log-RD, paraprotein, beta2, albumin, t(4;14) and gain(1q21) which were also important in predicting the OS. We again performed checks of the PH assumption for the first imputed dataset by looking at the Schoenfeld residuals which showed that the PH assumption was generally not

	Adjusted Log-HR (95% CI)	SE	p	FMI
TTR models				
Log-RD	0.107 (-0.125, 0.339)	0.118	0.365	0.147
Paraprotein	0.253 (0.133, 0.373)	0.061	< 0.001	0.239
Beta2	0.157 (0.019, 0.294)	0.069	0.026	0.367
Albumin	-0.074 (-0.193, 0.044)	0.060	0.216	0.112
t(4;14): Yes	1.163 (0.653, 1.674)	0.247	< 0.001	0.653
Gain(1q21): Yes	0.442 (0.129, 0.755)	0.156	0.006	0.427
OS models				
Log-RD	0.043 (-0.280, 0.365)	0.164	0.796	0.128
Paraprotein	0.183 (0.018, 0.347)	0.084	0.029	0.170
Beta2	0.247 (0.084, 0.409)	0.081	0.004	0.414
Albumin	-0.056 (-0.225, 0.113)	0.086	0.515	0.202
Hyperdiploidy: Yes	-0.442 (-0.844, -0.039)	0.203	0.032	0.280
t(4;14): Yes	0.489 (-0.079, 1.057)	0.280	0.090	0.526
Gain(1q21): Yes	0.493 (0.077, 0.909)	0.208	0.021	0.387

Table 3.18: Adjusted log-HR estimates and SEs, in brackets, from multivariable Cox PH models for TTR and OS in Myeloma fitted to covariates using MI in Stata to account for missing covariate data. The FMI and 95% CIs are also reported.

violated (plots not shown).

3.4.5 Increasing the number of imputations versus using Bayesian models to account for missing data in univariable Cox PH models for TTR and OS in Myeloma

The MI enabled us to make inference based on the full dataset, thereby having more power to detect covariate effects on the TTR and OS. We used FMIs and SEs to assess whether the MI led to improved estimates. To properly use MI, the recommended num-

ber of imputations, m, should depend on the proportion of missing data. In the previous analysis, we chose $m = 10$ as others have shown that 5-10 imputations would normally suffice and that there is no added benefit to increasing the increasing the number of imputations [121]. We increased the number of imputed datasets to 40 and again fit univariable Cox PH models for TTR and OS to assess if increasing the number of imputations led to more precise estimates. For comparison, we also fit Bayesian Cox PH models that account for the missing data by treating them as parameters for which we seek to obtain the posterior distribution alongside the log-HR. In this way, we were able to specify a joint model for the observed and missing data including the model parameters in order to proceed with the estimation in MCMC. Within the Bayesian models, we specified a linear model for continuous covariates with missing data, and logistic regression models for binary variables and then specified prior distributions for the parameters in these models. While we used the auxiliary variables age, sex, censoring indicator, log-transformed TTR or OS alongside log-RD and the other covariates in the MI, we did not include covariates in the models for missing data in the Bayesian models. We present estimates of the combined log-HRs (95% CIs) according to Rubin's rules, their SEs and FMIs from the models fitted to the 40 imputed datasets, as well log-HRs (95% CrIs) and SEs from the Bayesian models for both TTR and OS, Table 3.19.

In the TTR models, similar log-HR estimates for beta2 and albumin were obtained in both the MI and Bayesian models. Increasing the number of imputations to 40 led to a slight decrease in the SE from 0.068 and 0.052 to 0.062 and 0.052 for beta2 and albumin respectively. Again compared to the results from 10 imputations in Table 3.17, the SEs for hyperdiploidy, gain(1q21) and paraprotein were bigger after the 40 imputations. On top of this, the log-HR for paraprotein was 0.180 after the 40 imputations while it was 0.200 after 10 imputations while the effect of t(11;14) on the TTR reduced from -0.377 to -0.296. The FMI was reduced after the 40 imputations for t(4;14) and t(11;14) from 0.739 and 0.600 to 0.604 and 0.296 respectively, but not for the other co-

variates where the FMI was bigger, implying increasing the number of imputations did not necessarily lead to massive improvements in the precision of the estimates. The Bayesian model on the hand gave robust estimates with small SEs in general with estimates similar to those from MI. Hyperdiploidy had an association with the TTR (CrI did not include a 0), an association not previously seen even after MI. In the OS models, the only improvements based on the FMI and smaller SEs were seen for beta2, t(4;14) and t(11;14) but not the other covariates. Nonetheless, the overall estimates did not differ much from those based on 10 imputations. Again, the Bayesian model produced similar estimates to those from MI. On top of this, the CrIs in the Bayesian models were similar and at times, narrower than the CIs from MI. The Bayesian models converged without problems when fitted to TTR and OS based on trace plots showing good mixing and density plots centred around a unique mean posterior mean of the estimated log-HR for selected covariates, and also when using the linear and logistic regression respectively for missing continuous and dichotomous covariates, Figure 3.15.

Based on this analysis, we can surmise that the 10 imputations were sufficient to account for the missing data in the Myeloma dataset. The increase in the FMI and SE for some of the covariates may well have been due to chance, or it could have pointed to a need to investigate the missing data mechanism further. As the focus of this analysis was to broadly identify important factors associated with TTR and OS in the Myeloma dataset, we did not investigate the mechanism further, limiting our analyses to the MAR assumption. The brief use of the Bayesian models to handle missing data illustrated the power of these methods in general even when no covariates were used in the predictive models for the missing observations, the Bayesian model estimates were similar to those obtained via MI and had smaller SEs in general. Specifying covariates on the missing observations would therefore further improve the estimates. We next briefly use MI to deal with missing data in the CLL dataset.

	Cox PH model			**Bayesian Cox PH model**	
	Log-HR (95% CI)	SE	FMI	Log-HR (95% CrI)	SE
TTR models					
Paraprotein	0.180 (0.068, 0.293)	0.057	0.222	0.197 (0.075, 0.298)	0.002
Beta2	0.196 (0.075, 0.317)	0.062	0.292	0.196 (0.085, 0.312)	0.002
Albumin	-0.138 (-0.241, -0.035)	0.052	0.001	-0.135 (-0.231, -0.035)	0.001
Hyperdiploidy: Yes	-0.239 (-0.539, 0.062)	0.152	0.467	-0.479 (-0.753, -0.207)	0.004
t(4;14): Yes	1.133 (0.674, 1.591)	0.231	0.604	1.191 (0.779, 1.557)	0.006
t(11;14): Yes	-0.296 (-0.647, 0.056)	0.179	0.296	-0.358 (-0.757, 0.019)	0.006
Gain(1q21): Yes	0.557 (0.211, 0.903)	0.175	0.576	0.407 (0.111, 0.721)	0.005
OS models					
Paraprotein	0.076 (-0.084, 0.235)	0.081	0.214	0.092 (-0.068, 0.242)	0.002
Beta2	0.293 (0.166, 0.420)	0.065	0.138	0.263 (0.118, 0.383)	0.002
Albumin	-0.156 (-0.298, -0.014)	0.073	0.002	-0.146 (-0.281, -0.011)	0.002
Hyperdiploidy: Yes	-0.660 (-1.063, -0.257)	0.204	0.408	-0.838 (-1.271, -0.465)	0.006
t(4;14): Yes	0.805 (0.351, 1.259)	0.231	0.384	0.742 (0.290, 1.228)	0.007
t(11;14): Yes	-0.356 (-0.926, 0.214)	0.289	0.435	-0.473 (-1.120, 0.096)	0.009
Gain(1q21): Yes	0.687 (0.271, 1.103)	0.211	0.475	0.488 (0.107, 0.852)	0.005

Table 3.19: Log-HR (95% CI) estimates and SEs as well as the FMI from univariable Cox PH models for TTR and OS in Myeloma fitted to covariates using MI in Stata to account for missing covariate data based on $m = 40$ imputations. The estimated log-HRs (95% CrI) and SEs are also reported from Bayesian Cox PH models with imputation models for missing data for each covariate.

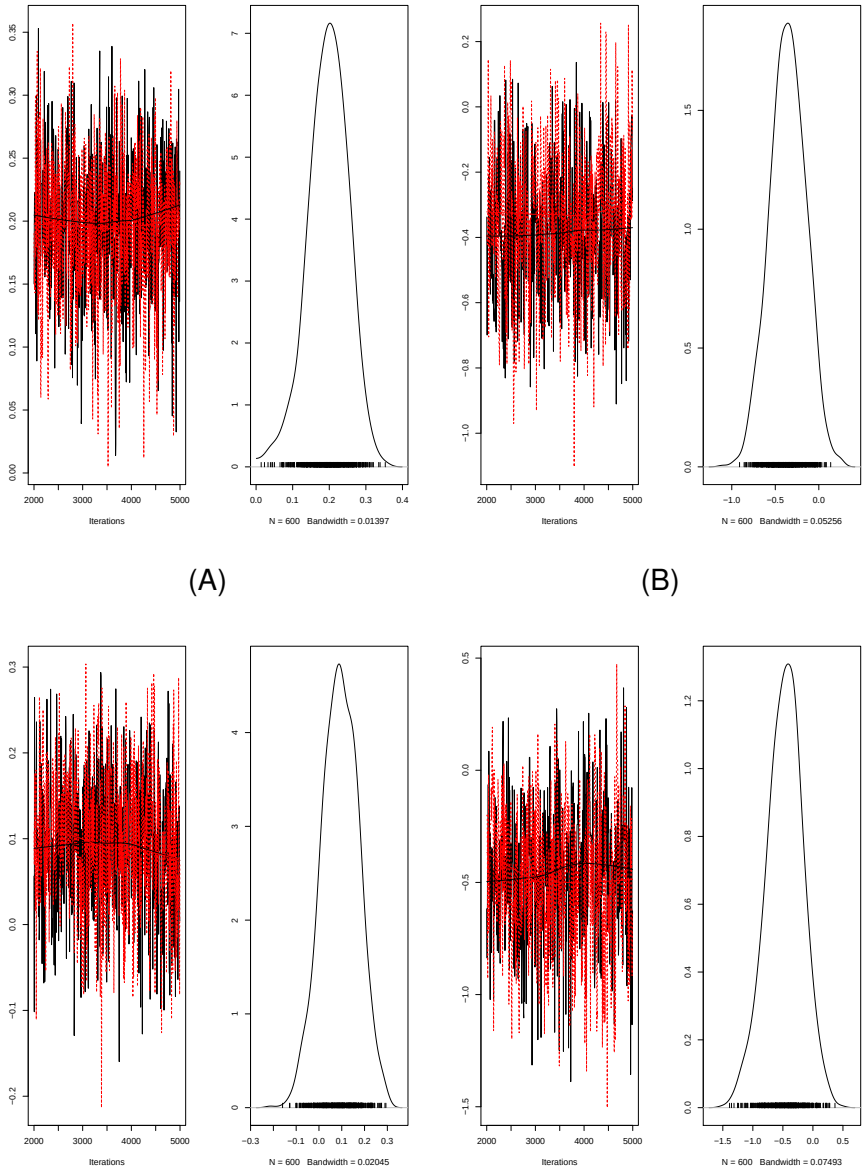


Figure 3.15: Trace and density plots for the coefficient of paraprotein (A) and t(11;14) (B) in Cox PH models for TTR. (C) and (D) are trace and density plots of the coefficients of paraprotein and t(11;14) respectively in Cox PH models for OS.

3.4.6 Missing data in the CLL dataset

Applying the same methods as before, we used MI to account for the missingness in the CLL dataset. Two covariates, VH mutation and p-deletion had missing data in the

CLL dataset, Table 3.9. A look at the missing data patterns showed that there was no systematic way in which data was missing for these two covariates, Table 3.20. There was complete data for 83% (271/325) of the patients and there were just 3 missing patterns for the 2 covariates each of which was observed less than 10% of the time. We again checked if the missingness in VH mutation and p-deletion was related to the other covariates; trial treatment, stage sex, age, log-RD, whether censored or not and TTR, using a logistic model with (1 if value missing and 0 if not missing) and found that the there was no evidence that missingness was not related to the other fully observed covariates (results not shown). We thus proceed to perform MI using MICE in Stata based on the logistic imputation model. From the resulting 10 imputed datasets, we compared the proportions of those with VH mutation and p-deletion between those had complete data and those whose values were imputed, From the resulting 10 imputed datasets, we compared the proportions of those with VH mutation and p-deletion between those had complete data and those whose values were imputed, Table 3.21. The proportion with VH mutation was higher among the complete cases compared to the incomplete cases. In a similar manner, the proportion with p-deletion was almost equivalent for the complete cases 96% and the incomplete cases 95.8%.

Number of patients	Percent	VH mutation	P-deletion
271	83	+	+
30	9	-	+
15	5	+	-
9	3	-	-
325	100		

Table 3.20: Missingness pattern in the CLL dataset. The (+) represents observed data while the (-) represents missing data.

Variable	n_{cc}	Complete cases	n_{ic}	Incomplete cases
VH mutation (Yes), fraction (%)	286	108 (37.8)	39	39 (36.2)
P-deletion (Yes), fraction (%)	301	289 (96.0)	24	23 (95.8)

Table 3.21: Proportion of those who had hyperdiploidy and p-deletion in CLL among the CCs compared to proportions after MI. n_{cc} represents number of complete cases while n_{ic} is the number of incomplete cases.

3.4.7 Univariable Cox PH models for TTR and OS in CLL fitted to imputed data

We then fit univariable models for TTR and OS to the imputed datasets and compared results to those from the CC models, Table 3.22. In the TTR models, having a VH mutation lowered the hazard of relapsing. The estimated log-HR was slightly bigger in the model fitted to the imputed data compared to the CC model. The SE was however smaller in this model based on the full data suggesting a more precise estimate. Moreover the FMI of 0.069 implied that only 7% of the total variability in this model was attributable to the missing data. For p-deletion, the effect was again bigger when modelling the TTR although it was not significant in either model. The SE was bigger in the imputed dataset.

For OS, the effect of VH mutation was significant in the CC model while it had an

	CC models		Models with MI		
	Log-HR (SE)	p	Log-HR (SE)	p	FMI
TTR model					
VH mutation: Yes	-0.626 (0.220)	0.004	-0.633 (0.217)	0.004	0.069
P-deletion: Yes	-0.620 (0.424)	0.144	-0.633 (0.444)	0.154	0.124
OS models					
VH mutation: Yes	0.203 (0.055)	< 0.001	-0.164 (0.306)	0.592	0.079
P-deletion: Yes	-0.252 (0.145)	0.082	-1.163 (0.537)	0.030	0.068

Table 3.22: Log-HR estimates and SEs, in brackets, from CC univariable Cox PH models fitted to covariates with missing data in the CLL dataset and equivalent results obtained via MI in Stata including the fraction of missing information (FMI) for each covariate.

opposite but non-significant effect on OS in the model fitted to imputed data. For p-deletion, the log-HR was five-fold larger for the full data model . This could be purely due to chance; since 96% of the patients had p-deletion, any associations between this variable and outcome are unlikely to be robust. The high proportion of patients with p-deletion may have been increased by the imputation since there would be little chance of sampling a p-deletion negative patient unless the number of imputations were to be increased substantially. Small changes in the data will result in large changes in the log-HR estimates. We thus did not include p-deletion in the multivariable models for this reason. Since VH mutation was not associated with OS, the final multivariable for OS in CLL would only include log-RD and stage which was already shown in Table 3.13.

We checked the validity of the PH assumption in all Cox PH models for TTR fitted to the 10 datasets and present residual plots from models fitted to the second imputed dataset, Figure 3.16. . We could not reject the PH assumption in the model with VH mutation as the p-value from the global test was not significant. For p-deletion on the

other hand, p was less than 0.05. As pointed out above, the very small number in the p-deletion negative group could give a higher probability of the imputed values falling in the group with p-deletion as seen in this plot. None of the imputed p-deletion values (black dots) fell in the smaller group which only had the observed values. A similar plot was obtained for OS but has not been shown.

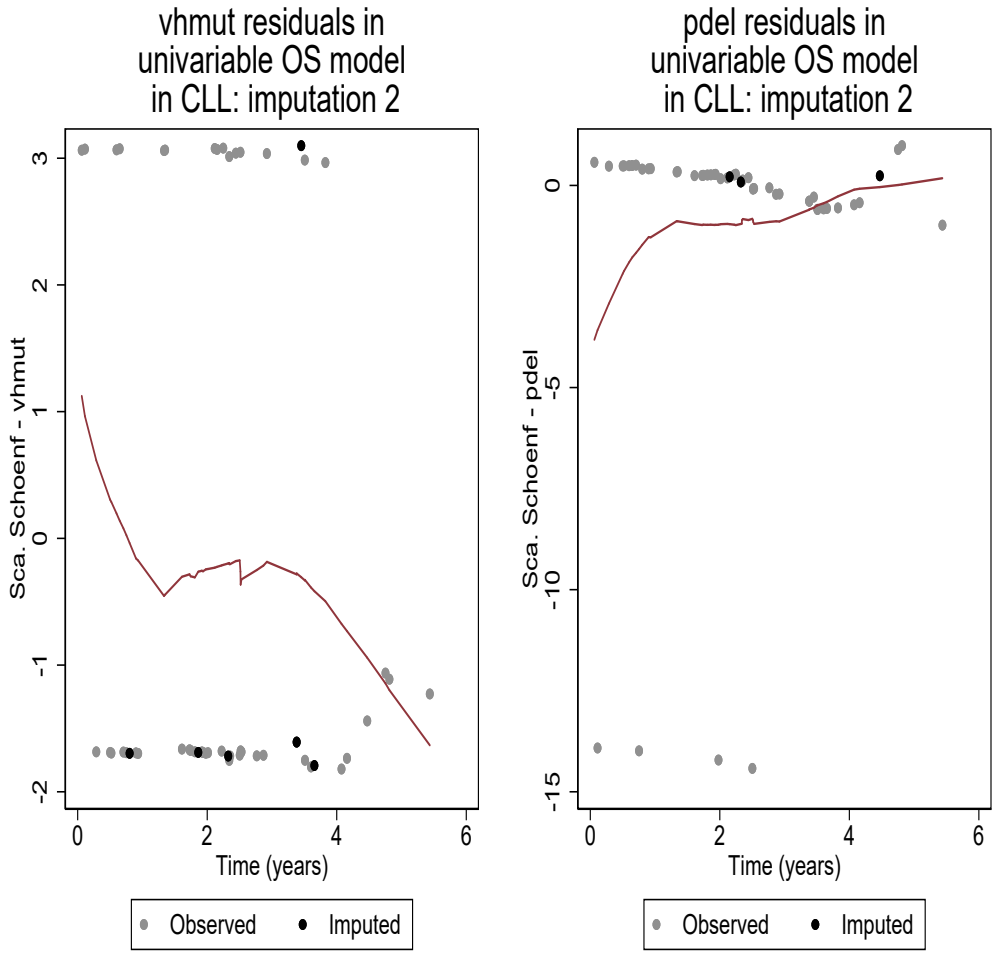


Figure 3.16: Schoenfeld residuals from univariable Cox PH models for TTR with VH mutation (vhmut), and p-deletion (pdel) as covariates, plotted against time since RD measurement using the second imputated dataset in CLL

	CC models		Models with MI		
	Adj. Log-HR (SE)	p	Adj. Log-HR (SE)	p	FMI
TTR models					
Log-RD	0.849 (0.083)	< 0.001	0.859 (0.080)	< 0.001	0.004
Stage C	-0.282 (0.210)	0.180	-0.261 (0.202)	0.196	0.005
VH mutation: Yes	-0.651 (0.224)	0.004	-0.646 (0.218)	0.003	0.050

Table 3.23: Adjusted log-HR estimates and SEs, in brackets, from multivariable Cox PH models for TTR in CLL fitted to covariates using MI in Stata to account for missing covariate data. The FMI and 95% CIs are also reported.

3.4.8 Multivariable Cox PH models for TTR in CLL fitted to imputed data

The same approach for selecting the final model was taken, after excluding p-deletion. As the variables in the final CC and model with MI were the same, we present the estimates from both models side by side in Table 3.22. Log-RD was associated with TTR in both models. The effect was very similar in both CC and full dataset analyses. The SE in the imputed dataset was slightly smaller indicating better precision, whilst the total variability attributed to the missing data was 0.4%. The log-HR for stage slightly smaller in the model based on imputed data where again the SE was smaller and the FMI was 0.005. There were no significant differences in the risk of a relapse by stage in either models. Finally, having VH mutation was predictive of longer TTR. Again the SE was smaller indicating more precise estimates resulting from the MI.

These results indicate that MI helped us obtain more precise estimates in the CLL dataset. By accounting for missing data, we were able to obtain estimates with smaller SEs which means the imputation achieved the desired goal. We therefore did not consider increasing the number of imputations or other means of modelling the missing data and we proceed to discuss the results from the various methods in context.

3.5 Summary

In this chapter, we explored factors influencing TTR and OS in Myeloma and CLL using standard TTE models as a way of introducing these datasets which will be used to explore how well various methods proposed in this thesis perform. We focused on log-RD as an indicator of the RD burden after treatment in these analyses, and other prognostic factors that are known to be associated with TTR and OS and are at the same time markers of the tumour growth rate which is not observed in practice.

Residual disease is clearly an important factor in the time to relapse or progression of Myeloma, and in the subsequent time to death. Therefore, treatments or combinations of treatments, that aim to eradicate as much of the disease as possible should be pursued. This is apparent even in this relatively small dataset. There is also some evidence that biomarkers for tumour growth, such as paraprotein, beta2 and albumin are important predictors disease progression, although the evidence is less clear. This may be expected since these variables are markers of growth, rather than direct evidence, which could only be studied if repeat measurements of tumour burden were available. It is likely that there is some association between RD and markers of tumour growth, since the coefficients were attenuated in the multivariable model. Translocations such as t(4;14) and gain(1q21) which have both been associated with TTR and OS in Myeloma, were also found to be significantly associated with these outcomes in this analysis [122].

This analysis also showed that log-RD is also an important predictor of both TTR and OS among the CLL patients, again confirming the need for treatments that are able to completely eradicate the disease where possible. For patients who have some RD, additional treatments may be required to prevent progression and subsequent death with CLL. We did not find evidence of a relationship between outcomes and disease stage, which is a measure of how aggressive the disease is. This is surprising and

may result from the low power in this data set and bias or imprecision due to missing data. VH mutation was an important, though not significant predictor of TTR, whilst p-deletion had a protective effect on the OS. VH mutation status has been known to correlate with the clinical outcomes in CLL although why those with VH mutation have better outcomes than those without this mutation actually depends on the proliferation rate of CLL cells, [123]. Similarly, patients with p-deletion have been known to not respond well to therapy, however, others have been known to remain relapse-free for a long time following treatment, which suggests that there could be other mechanisms that drive relapse and time to death in CLL which cannot be teased out in this descriptive analysis, [124].

Both Exponential and Weibull parametric models were explored in the analysis, with the shape parameters for the Weibull model significantly bigger than 1 models for OS in Myeloma and both TTR and OS models in CLL. This suggests that the baseline hazard is increasing over time rather than constant. The estimated log-hazard ratios were similar in the Cox PH and Weibull models, so that if interest is only on the the effect of the variables on the TTE outcomes, either of the two models could be used to good effect for these datasets. However, as interest is usually not on the underlying baseline hazard, it would suffice to use the Cox PH model.

The estimated survivor curve especially for the TTR following RD measurement in the Myeloma dataset may suggest the presence of a 'plateau' at the tail end of the distribution of the relapse times. This could mean that not all patients would relapse after treatment and therefore using the standard TTR or OS models ignores the existence of this proportion of patients. We thus used cure rate models to investigate if there was a proportion cured and discuss the application of these methods in the next chapter.

Chapter 4

Estimating the cured proportion following treatment in cancer TTE models

4.1 Introduction

In this chapter, we examine methods for estimating the proportion that is cured of the disease following treatment using the two definitions of cure in Chapter 1 and using the cure rate models introduced in Section 2.6.

TTE methods assume all patients will relapse or die following treatment. However, in some cancers, there is a proportion (π) of individuals that will never experience the event following treatment and are therefore considered 'cured' of the malignancy. The proportion cured or the cure rate is an important measure of long-term survival benefit. Plots of the estimated survival functions from a group of patients where a proportion is cured will reach a non-zero asymptote indicating a cure proportion. Standard analysis methods are thus inappropriate in this case as they assume that all individuals will have the event given sufficiently long follow-up.

In this thesis, we consider two types of cure rate models. The first type of cure rate models assume that the population of interest contains two distinct groups of individ-

uals; those who are considered cured of the disease following treatment and are no longer at a risk of a relapse while the rest who are not cured, are still prone to a relapse following treatment [18]. The second type of cure rate models we consider are part of a broader class of models referred to as relative survival models [69]. These models are used for estimating the proportion of patients whose OS is equivalent to that of age-sex matched individuals from the general population without the disease.

The cure rate models are broadly classed as mixture or non-mixture models with a parametric or semi-parametric model specified for the TTR or OS for those that are not cured . We briefly discuss each of these models next.

4.2 Parametric mixture models

Mixture models work on the premise that there are 2 groups of patients treated for a particular disease. The first group consists of patients who are cured while the rest includes the rest of the patients who are not cured by the treatment, hence the name mixture models. To segregate those who are cured from the rest, a logistic model is normally used [62], while a standard TTE model such a Cox PH, Weibull and others can be used to model the time to death or TTR for those who are not cured [62, 26]. Covariates can thus be used to investigate both the probability of being cured and the TTR/OS for those who are not cured. This makes it easier investigate the influence of covariates on the cured proportion and the TTR and/or OS for those who are not cured separately. This means a covariate effect can be significant on the probability of being cured and not on the TTR/OS among those not cured and vice versa. Mixture models are therefore convenient where interest is in ascertaining whether there is a proportion who are cured following treatment while at the same time modelling the TTR or OS for the rest. Mixture models are ideal for investigating whether there exists a proportion of patients who might be cured by the treatment and are therefore the mostly commonly

used cure rate models.

In this thesis, we use mixture models to estimate the proportion, π, of patients who will never relapse following treatment and specify a parametric distribution for the TTR for those who are not cured following treatment [125] in the Myeloma and CLL datasets. Using the formulation in Section 2.6. Based on data from n individuals, we can express the full likelihood function following (2.5) as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} (1-\pi_i)^{\delta_i} [f_r(t_i)]^{\delta_i} [\pi_i + (1-\pi_i) S_r(t_i)]^{1-\delta_i} \tag{4.1}$$

where both π_i and t_i are conditional on some observed covariates and $\boldsymbol{\theta}$ is a vector of parameters on π and the TTR distribution for those not cured. The models are fitted by specifying a pdf, $f_r(t_i)$ with associated survivor function $S_r(t_i)$, for those who will relapse, which might take any of the known parametric forms for non-negative random variables such as the Weibull, Gamma or log-normal distributions. In this thesis, the logistic model for π_i was specified as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1 z_{1i} + \cdots + \alpha_q z_{qi} \tag{4.2}$$

where the i-th individual has q observed covariates and the α's represent the coefficients quantifying the association of each covariate with π_i and α_0 represents the intercept. For those not cured, covariates can be modelled through one of the parameters governing the TTR distribution. For this work, covariates on the TTR for those not cured were modelled through the scale parameter λ,

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

where the β's represent the association of each of the p observed covariates with the TTR.

4.3 Parametric non-mixture models

An alternative cure rate model used for estimating the proportion that will never relapse following treatment is the non-mixture model which is also called the promotion time cure (PTC) model. The PTC model estimates the cured proportion by defining an asymptote for the cumulative hazard function. In this model, it is assumed that treatment will leave a number, N, of cancer cells which will, after some time, re-grow to a detectable level at which point the patient is said to have relapsed [126]. A Poisson distribution with mean ρ is assumed for this number of remaining cancerous cells. If M_j is a random variable representing the time it takes for the j-th cancer cell to grow back to detectable disease, then the TTR is the time it will take for any of the cancerous cells to grow back to detectable level $T = min(M_j, j = 1, 2, ..., N)$. And assuming the M_j are independently and identically distributed with a given distribution, the survivor function for T can be given as

$$\begin{aligned} S(t) &= Pr[\text{No cancer cells at time t}] \\ &= Pr[N = 0] + Pr[M_1 > t, M_2 > t, \ldots, M_N > t, N \geq 1] \\ &= e^{-\rho} + \textstyle\sum_{N=1}^{\infty} S^N(t) \frac{\rho^N - \rho}{N!} \\ &= e^{-\rho(1-S(t))} = e^{-\rho F(t)} = \pi^{F(t)} \end{aligned} \tag{4.3}$$

where $\pi = e^{-\rho}$ represents the proportion that is cured and will therefore never relapse [127]. Further, note that

$$\lim_{t \to \infty} S(t) = \pi = e^{-\rho} > 0$$

such that $S(t)$ is an improper survivor function which does not tend to 0 as t gets very large. From (4.3), we can derive the hazard function as

$$h(t) = -(\log\pi) f_r(t). \tag{4.4}$$

Then, given a random sample of n individuals, we can write the likelihood function in terms of the hazard and survival functions as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [-(\log\pi_i) f_r(t_i)]^{\delta_i} \exp[F_r(t_i) \log(\pi_i)]. \tag{4.5}$$

As before, $f_r(.)$ and $S_r(.)$ may follow any of the standard parametric distributions for non-negative random variables. The effect of covariates on the cured proportion π, and the TTR for the chosen parametric distribution be modelled using the logistic model (4.2) and through the scale parameter for the chosen TTR distribution as in the mixture model in Section 4.2. Other link functions such as the log-log link have been used in the model for π in PTC models. This link function is normally used when interest is in obtaining covariate effects that are log-hazard ratios when the parameters within the distribution function for T do not vary by covariates [29]. In this way, the PTC model has a PH structure. In our analysis, we will use the logistic regression to model the effect of covariates on π as in the mixture models to enable comparison of estimates between the two models. Covariates on $f_r(.)$ and $S_r(.)$ will have a slightly different interpretation as short-term effects on T, while not necessarily characterising the TTR for those who are not cured [128]. To obtain the survivor function for those who are not cured, we can express (4.3) in a form similar to the mixture model formulation of equation (2.16) in Section 2.6 as

$$S(t_i) = \pi_i + (1 - \pi_i) \left(\frac{\pi_i^{F_r(t_i)} - \pi_i}{1 - \pi_i} \right) \tag{4.6}$$

which resembles the survivor function in the mixture model where the survivor function for those who are not cured is $\left(\frac{\pi_i^{Fr(t_i)}-\pi_i}{1-\pi_i}\right)$.

The PTC model attaches biological meaning to the analysis based on its formulation which makes it more attractive than the mixture model. Another important feature of the PTC model is that it can take on a PH model structure which makes it possible to focus on the effect of covariates on π. The other advantage is that it has a simple structure for the survival function which makes computation easier when compared to the mixture model [129]. The PTC models are therefore more useful where interest is in understanding the biology or mechanisms that drive outcomes while at the same time investigating the role of covariates on the probability of being cured.

4.4 Semi-parametric mixture cure models

The complexities with having to specify and then verify the suitability of a chosen parametric form for the survival function for those who will relapse have led some to propose the semi-parametric mixture cure (SPMC) model [25, 26]. For this model, the only difference is that the parametric form of the survivor function for modelling the TTR, given that it occurs, is replaced by the PH model. The survivor function can still be expressed as

$$S(t) = \pi + (1-\pi)S_r(t) \tag{4.7}$$

where $0 < \pi < 1$, is the probability of being cured. If $\boldsymbol{z}$ and $\boldsymbol{x}$ are vectors of covariates on the cure proportion and TTR respectively as previously described, and the PH assumption holds, then we can replace the survival function for those bound to relapse in (4.7), by the Cox PH model. We can thus write the survival function for the i-th individual as

$$S(t_i) = \pi_i + (1 - \pi_i) S_{0r}(t_i)^{\exp(\boldsymbol{\beta^T x_i})} \tag{4.8}$$

where π_i is related to the $\boldsymbol{z_i}$ through a logistic model (4.2) and $S_{0r}(t_i) = \exp(-H_{0r}(t_i))$ is an unspecified baseline survivor function for those who will relapse where $H_{0r}(t_i)$ is the cumulative hazard up to time t_i. These models can be fitted using the EM algorithm discussed in Section 2.9.2 in standard statistical software.

4.5 Cure rate models that incorporate population survival data

In instances where OS is the outcome of interest, relative survival models can be used to estimate the proportion of patients whose survival is similar to that of disease-free members of the general population. In these situations, 'cure' is observed when the hazard rate for death in individuals with a particular disease is the same as that of the general population. Estimating the cured proportion by incorporating information about the survival in the general population helps deal with selection bias which might result from estimating the cured proportion based on a single sample from a trial as in the previous section [130]. In this thesis, we refer to these models as population mixture and PTC models.

4.5.1 Population mixture models

At the population level, the overall survivor function is the product of the expected survival function in the general population without the disease $S_p(t)$ and the assumed survival function attributable to the disease $S_d(t)$ [130]. Using (2.17), the hazard function can be worked out as the sum of the population hazard $h_p(t)$ and that attributable to the disease and using the pdf for those bound to die from the disease to give the overall hazard function

$$h(t) = h_p(t) + \frac{(1-\pi)f_d(t)}{\pi + (1-\pi)f_d(t)}, \tag{4.9}$$

since $h(t) = f(t)/S(t)$. We can then use these hazard and survival functions to get the likelihood for n individuals. Estimates of the parameters can be obtained by maximising the resulting likelihood as long as information is available on the population hazard, $h_p(t)$, from population mortality statistics, for example. Again, given data from n individuals, the effect of covariates on the proportion whose OS returns to that of age-sex matched individuals in the general population and the OS for those who are not cured can be modelled using the logistic model specified in (4.2) and through the scale parameter (λ) if we assume a two-parameter parametric distribution such as the Weibull for the OS of those not cured.

4.5.2 Population non-mixture models

The population PTC model has the same motivation as the PTC models described for TTR. The survival function in (4.3) can be replaced by the overall survival function which is the product of the population survival function $S_p(t)$ and a function of the CDF for those that will eventually die from the disease $F_d(t)$ as in [69], as follows

$$S(t) = S_p(t)\pi^{F_d(t)} \tag{4.10}$$

and this can be written as

$$S(t) = S_p(t)\text{exp}(\text{log}(\pi)(1 - S_d(t)) \tag{4.11}$$

since $F_d(t) = 1 - S_d(t)$, while the hazard function can now be replaced with the sum of the two hazard functions, that for the general population $h_p(t)$, and $h_d(t)$ for those that will eventually die, that is

$$h(t) = h_p(t) - \log(\pi) f_d(t). \tag{4.12}$$

The LL, given n individuals, can similarly be derived from hazard and survival functions as before. The term $S_p(t)$ does not depend on the parameters from the distribution governing the survival of those not cured, $f_d(t)$, and can thus be ignored when estimating the parameters of the distribution.

4.6 Application of cure rate models to the Myeloma and CLL datasets

The various cure rate models discussed were used to estimate the proportion of patients who will never relapse and the proportion whose OS is the same as age-sex matched individuals in the general population in the Myeloma and CLL datasets. In all models, covariates on the cured proportion were modelled using logistic regression. Covariates associated with the TTR and OS were modelled through the scale parameter in Weibull models and using Cox PH models in semi-parametric mixture models. All Bayesian parametric models were fitted using the MCMC methods described in Section 2.9.3. To estimate the proportion, π, that will never relapse following treatment as well as model TTR for those who were not cured, three types of cure rate models were used:

1. A semi-parametric mixture model with a Cox PH model for those not cured estimated by ML, using the R package **smcure** [25]
2. A Bayesian parametric mixture model with a Weibull model for the TTR, and
3. A Bayesian parametric PTC model with a Weibull model for the TTR. The semi-parametric mixture model is well established and should therefore produce estimates similar to the Bayesian models if there exists a proportion that will never relapse following treatment.

The proportion whose OS will be similar to age-sex matched individuals in the general population was estimated using three types of cure rate models:

1. A Weibull model for those that will eventually die from the disease, estimated by ML using the Stata function **strsmix** [69],
2. A Bayesian parametric mixture model with a Weibull model for the OS, and
3. A Bayesian parametric PTC model with a Weibull model for the OS.

Prior to fitting these, the hazard rates, $h_p(t)$, in (4.9) and (4.12) we obtained from the ONS data as described in Section 2.7.

4.6.1 Prior distributions for model parameters in Bayesian cure rate models

To fit the Bayesian models, prior distributions were specified for both the parameters associated with the cured proportion and the parameters in the TTR and OS models for those not cured.

When estimating π without covariates, a $Uniform(0,1)$ prior was used for the cured proportion. In the Weibull model for those not cured and bound to relapse or die from the disease $Gamma(1,1)$ priors were chosen for both the scale (λ) and shape (γ) parameters.

For all other models with covariates, a vague $N(0,100)$ prior was used for both the intercepts and coefficients in both the logistic model for the cured proportion and the parameters in the models for TTR or OS for those not cured.

4.7 Cure rate models without covariates in Myeloma

First, the cured proportion π, representing the proportion that will not relapse, was estimated without covariates using the Bayesian mixture and PTC models. The semi-

parametric mixture model was not used to estimate π on its own as **smcure** does not allow for the estimation of the cured proportion without covariates on both π and the TTR to ensure identifiability [25].

Estimates of the cured proportion from the Bayesian mixture and PTC models are shown in Table 4.1. The estimated proportion who will never relapse following was just under 20% in both the mixture and PTC models. Estimates of π from the two models including the CrIs were similar. The shape parameter in the PTC model was bigger than in the mixture model suggesting a higher risk of relapsing for those not cured by treatment.

	Mixture model	**PTC model**
Parameter	Estimate (95% CrI)	Estimate (95% CrI)
Cure model		
π	0.197 (0.147, 0.246)	0.184 (0.126, 0.242)
TTR model		
λ	0.391 (0.330, 0.456)	0.201 (0.155, 0.247)
γ	1.290 (1.159, 1.424)	1.409 (1.260, 1.566)

Table 4.1: Estimates of π, the scale (λ) and shape (γ) parameters from Bayesian Weibull mixture and PTC models for TTR in the Myeloma dataset.

Finally, a plot of the TTR for the cured and non-cure subgroups combined (Overall TTR, blue line), estimated from the posterior distribution using MCMC according to equation (2.16), was similar to the K-M estimate for all patients, with the characteristic plateau observed at around 20%. Considering the non-cured patients only (red line), the TTR survivor curve is fully observed (all patients progressed) by the end of follow-up in both models, Figure 4.1.

Both models showed a good fit to the data with good convergence as seen in the trace, density and auto-correlation function (ACF) plots in Figure 4.2. The trace plots in both models showed good mixing, while the density plots for both π and λ were centred

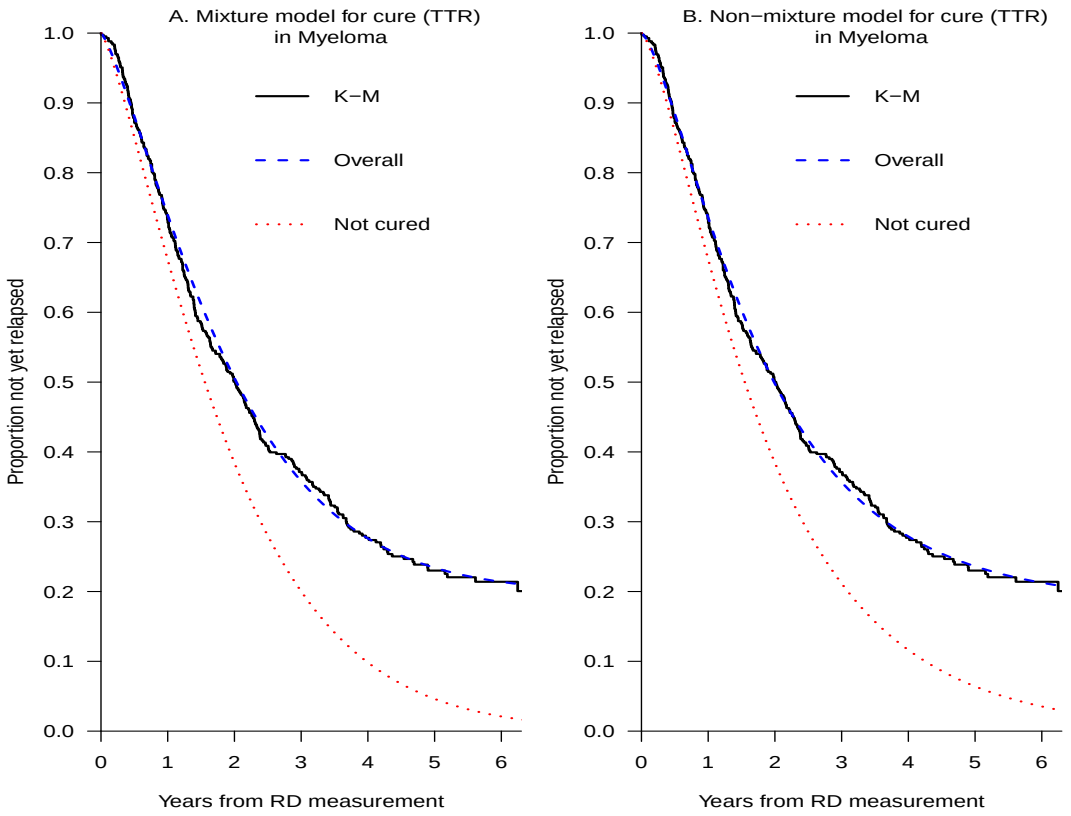
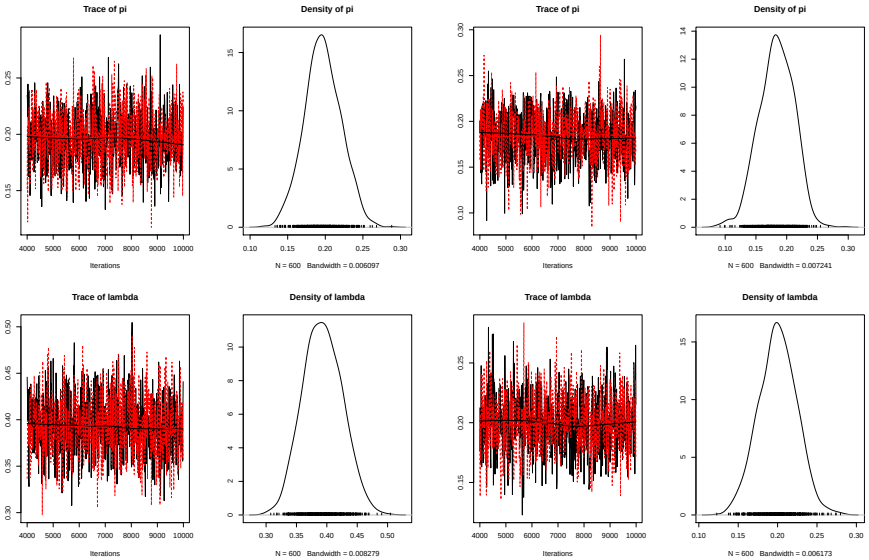


Figure 4.1: Model estimated and K-M survivor curves from Bayesian Weibull mixture (A) and PTC (B) models for TTR in the Myeloma dataset

around the posterior means.



(A) Bayesian mixture (B) Bayesian PTC

Figure 4.2: Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian Weibull mixture (A) and PTC (B) models for TTR fitted to the Myeloma dataset.

4.7.1 Cure rate models with covariates in Myeloma

The influence of log-RD as well as z-standardised paraprotein, beta2 and albumin as independent covariates on both the cured proportion and the TTR for those not cured in the mixture models were simultaneously estimated from the Myeloma data using the models described in Section 4.6. Log-odds ratios (ORs) were obtained from the logistical model for the cured proportion, while log-HRs were estimated from the models for TTR. On top of the log-HRs, the shape parameter and intercept were estimated in the Bayesian Weibull mixture and PTC models.

The semi-parametric mixture model was fitted using the smcure package in R, while the Bayesian Weibull and PTC models were fitted using MCMC in OpenBUGS. Albumin and beta2 were again used to give a measure for disease stage while sex and age were not included at all as they were not significantly associated with TTR as shown

previously. The estimates from these models are shown in Table 4.2.

An increase in the log-RD was significantly associated with a lower probability of being cured following treatment. Each log-increase in the RD percentage after treatment was predictive of a decrease in the log-OR of -0.804 (-1.455, -0.154) in the PH mixture model, -0.739 (-1.228, -0.206) in the Weibull mixture model and -0.791 (-1.457, -0.152) in the Weibull PTC model respectively. The log-RD was however, not significantly associated with TTR in any of the three models. This is in contrast to the standard univariable TTR models where a high log-RD was predictive of a shorter TTR. Actually, log-RD had a protective effect in the Bayesian mixture and PTC models although the CrIs in these models included a 0. The estimated log-HR was biggest in absolute value in the PTC model. The DIC in the two Bayesian models was not very different for log-RD.

Paraprotein is one of the known markers of disease burden, with high levels of paraprotein associated with a shorter TTR. We thus fitted paraprotein as a covariate on both the cured proportion and the TTR for those not cured in the mixture models, and to model the short-term effect of paraprotein on the TTR in the PTC model, considering only those patients who had data on paraprotein ($n = 334$) in a CC analysis. In all three models, higher levels of paraprotein were associated with a lower probability of being cured, log-OR -0.668 in the PH mixture model, -0.621 in the Bayesian Weibull mixture model and -0.730 in the Bayesian Weibull PTC model respectively. However, the 95% CI in the PH mixture and CrIs in the Bayesian equivalents showed that this effect was not significant. For the TTR, a bigger level of paraprotein was associated with a higher risk of relapsing and the effect was significant in the PH mixture model log-HR (95% CI) 0.203 (0.035, 0.372) but not in the Bayesian mixture and PTC models.

A higher value of beta2 was associated with a shorter TTR and a lower probability of being cured in all the models but this was not significant. Log-OR estimates were variable across the 3 models. For albumin, higher levels of were associated with a lower probability of being cured (positive log-ORs) in the cure models and lower chances of

relapsing (negative log-HRs) in the TTR models for those not cured and bound to relapse. Again the models with beta2 and albumin had respectively 310 and 426 patients in a CC analysis ignoring the missing data.

These results show that while it is possible to obtain the cured proportion from both the mixture model and the PTC model, the reliability of the results especially in the PTC model can be put in question. This may be due to the relatively small sample size and the complex nature of the statistical models. It was also not possible to add multiple covariates due to the small sample size. To properly model this, we need to take into account the nature of the data i.e. presence of a plateau that indicates that some patients have indeed been cured.

On comparison, the mixture model generally performed better than the PH mixture and the PTC models as it had shorter CrIs in general and therefore more precise estimates. Looking the DIC, the PTC model seemed to perform better than the mixture model. However, the reduction in the DIC was not big to signify it was the better model. All models reported estimates of log-ORs in the same direction. Differences were seen when it came to predict the TTR having taken into account the proportion who might have been cured from the disease where the mixture model performed better.

Comparing TTR patterns by log-RD level

We briefly investigated whether it was reasonable to assume the existence of a proportion who might have been cured among the Myeloma patients. Using the estimated survivor function from the PH mixture model, we plotted predicted values for the overall survivor function assuming log-RD values of -2, -1, 0, 1 and 2 respectively representing RD percentage values of 0.1%, 0.4%, 1%, 2.7% and 7.4%, which in the context of the available Myeloma data represented a high disease burden, Figure 4.3. We also plotted the K-M curve for all patients as well as the estimated cured proportion ($\hat{\pi} = 0.197$) from the Bayesian mixture model since smcure does not fit PH mixture models without covariates on π, alongside the lower and upper limits of the 95% CrI. It is worth noting that the K-M survivor curve for all patients approaches an asymptote at the estimated cured proportion from the mixture model. This means it is reasonable to use cure rate models when characterising the TTR in the Myeloma dataset.

A log-decrease in the RD percentage from -1 to -2, 0.4% to 0.1% on the original scale, increase the probability of being cured from around 20 - 36%, figure on the left. For those with a log-RD of 0, representing a 1% RD percentage following treatment, the estimated survivor function plateaued at around 10% meaning 90% of individuals in this group would relapse while the remaining 10% could be cured. Another log-increase in the RD percentage to 2.7% reduced the cured proportion from 10 to 5% and finally, only 2% of those with a high disease burden representing a log-RD of -2, had a chance of not relapsing again after treatment.

These findings illustrate that the cured proportion decreased rapidly with increasing log-RD, emphasising the effect of log-RD in all the three models. The coinciding of the K-M survivor function and the estimated cured proportion also confirmed that we can estimate the proportion who might have been cured and therefore no longer a risk of relapsing in the Myeloma dataset.

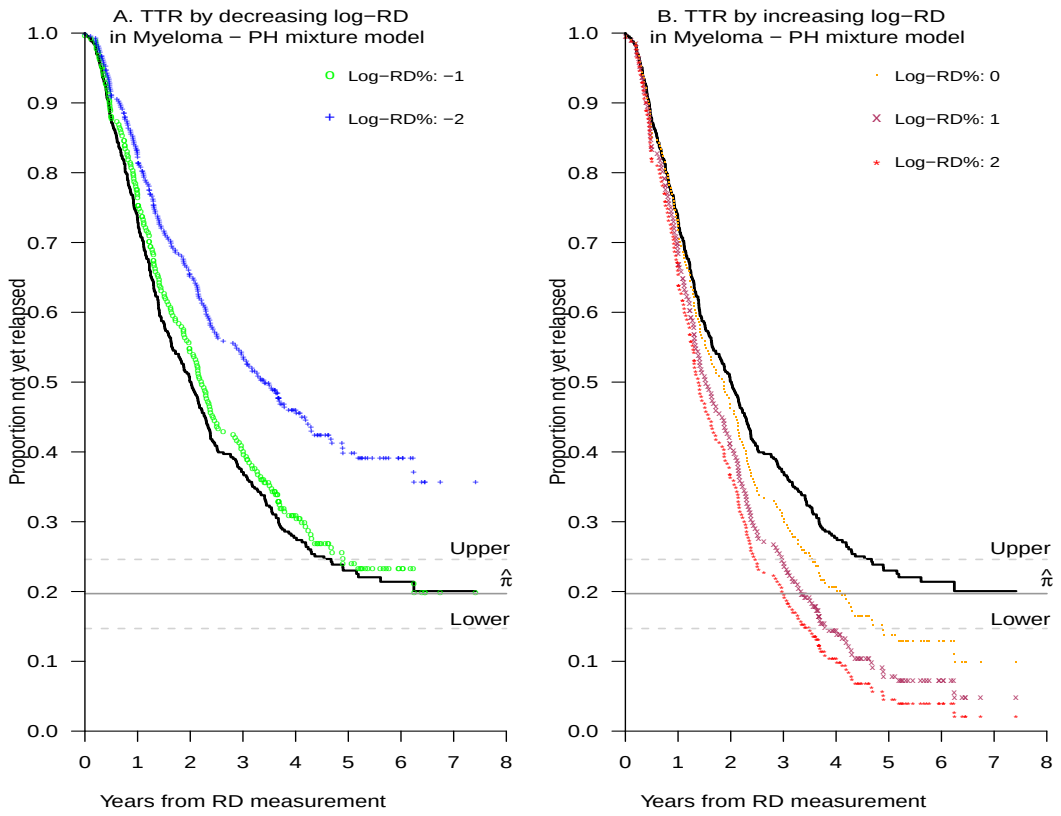


Figure 4.3: Model estimated survivor functions by selected log-RD values from the PH mixture model fitted using smcure to the Myeloma data. Graph (A) shows curves with decreasing log-RD while (B) are curves with increasing log-RD. The grey horizontal line shows the estimated π from the Bayesian mixture model while dotted horizontal lines are upper and lower limits of the 95% CrI for the cured proportion. The solid black line represents the K-M estimate of the survivor function for all patients.

4.8 Estimating the proportion whose OS returns to that of general population in Myeloma

To incorporate population survival data, population mixture model and PTC model were fitted to the Myeloma data to estimate the proportion whose OS returned to that of age-

sex matched members of the general population and model OS for those not cured. The mixture models were fitted using ML in the Stata package **strsmix**, [69], and using Bayesian methods implemented in OpenBUGS. All the PTC models were fitted in OpenBUGS.

The estimates (95% CI or CrIs) of π from Weibull ML, Bayesian mixture model and the Bayesian PTC model were 31%, 28%, (2 - 51 %) and 29%, (3 - 50%) respectively, Table 4.3. This is approximately 10% more than the estimated proportion who would never relapse again in the previous section, which suggests that 10% of cases may relapse but die of something other than Myeloma. The uncertainty intervals for the estimates of π from all 3 models were much wider than those in the TTR models. Around 30% of these patients would return to the same survival distribution as age-sex matched controls in the general population after the treatment. Plotted survivor curves from the Bayesian mixture model and the Bayesian PTC model showed that the estimated overall survival function was roughly close to the empirical survival curve (blue line), while the survival function for those not cured rapidly decreased towards 0 as $t \to \infty$ in the mixture model, Figure 4.4. It is worth noting that in both models, the estimated overall survivor curves did not decrease in such a manner to suggest the curves were approaching a non-zero asymptote at which cure was possible. With OS in Myeloma, a longer follow-up would be needed to observe this asymptote if it exists.

The trace and density plots from the Bayesian mixture and PTC models, Figure 4.5 showed that in general the models for π without covariates converged to the target posterior distribution. However, the density plots from the mixture and PTC models clearly showed that no unique value for the cured proportion was identified which could have led to the wide CrIs. We thus could not establish whether for these patients, there was a proportion whose OS might have returned to the population OS after treatment.

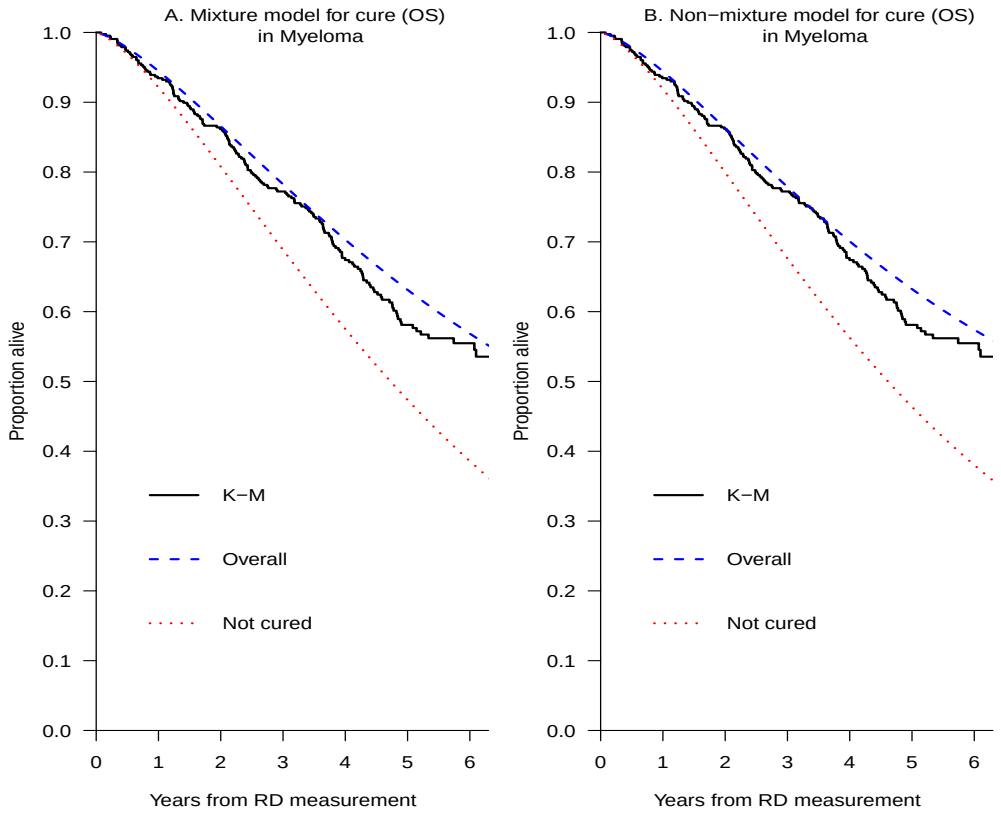


Figure 4.4: Estimated K-M curve and overall survivor (blue dotted line) curves and OS curves for those not cured (red dotted line) from the Bayesian Weibull population mixture (A) and Bayesian population PTC (B) models respectively.

4.8.1 Modelling covariates in cure rate models for OS in Myeloma

To investigate the role of covariates on the proportion of patients whose survival is similar to disease-free individuals in the general population, we again fit the three models with the same covariates on both the cured proportion π and the scale parameter λ of

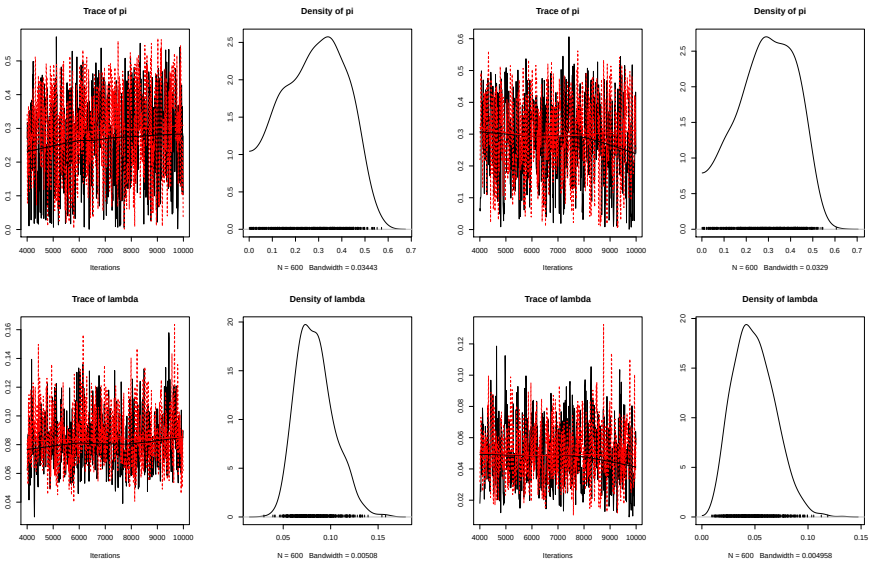


Figure 4.5: Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian population Weibull mixture (A) and Bayesian Weibull PTC (B) models for OS fitted to the Myeloma dataset.

the Weibull models while incorporating age-sex matched population hazard and survival rates for each individual in the data set from the UK ONS. The results are shown in Table 4.4.

There was no significant association between log-RD and the probability of having OS similar to disease-free individuals in all three models. The point estimates in the ML based Weibull mixture and the Bayesian PTC models indicate that those with lower log-RD after treatment would have a higher probability of not dying from the lower disease log-OR -0.313 in the ML mixture, -0.471 in the Bayesian mixture and 0.138 in the PTC model. The estimate in the PTC model implied a higher log-RD predicted a higher chance of attaining population OS. However, all effects of log-RD were not significant as the CI and CrIs included a 0. While the models did not predict the probability of being attaining population OS, they did not also predict the time to death either. The ML model estimate actually showed a protective effect of log-RD on the OS, (log-HD

-0.055).

None of the three models showed a discernible effect of paraprotein on the cure probability. In the Bayesian mixture and PTC models for instance, higher levels of paraprotein were predictive of a higher chance of the OS returning to that of disease free individuals while in the ML mixture model, lower levels of paraprotein led to a higher chance of attaining the population OS. The point estimates of the log-OR from all models differed greatly with the largest effect seen in the Bayesian mixture model. The CIs and CrIs confirmed that the effect of paraprotein on the proportion could not be precisely estimated. The trace and density plot for the log-OR (alpha1) for paraprotein in the logistic model for π and the log-HR (beta1) for paraprotein in the OS model are shown in Figure 4.6. The figures illustrated that while convergence might have been attained, the effect on the cured proportion could no be precisely estimated as seen from the density plot and also from the wide CrI. The effect of paraprotein on OS was not significant in all models.

With respect to beta2, lower measures were suggestive of a higher probability of not dying from the disease in all models. The estimates of the log-ORs in the Bayesian mixture and PTC models were larger than those from the ML mixture with very wide CrIs which implied beta2 was not a predictor for π in this dataset. For the OS, beta2 was significantly associated with the time to death in the ML mixture model but not in the other two models although the direction of the effect was the same.

Finally, albumin was not associated with the probability of attaining the population OS in any of the models. The point estimates suggested lower albumin would increase the chances of attaining the population OS in all three models. However, these estimates had wide CIs or CrIs. Moreover, it is worth noting that throughout the previous analyses, a higher level of albumin was indicative of adverse outcomes meaning these effects should have been negative if albumin was a predictor for π. On the other hand, albumin was associated with OS with higher values of albumin at baseline predictive of

longer time to death. In the OS model, the expected association of albumin was seen.

Including covariates in the population cure rate models for OS did not provide any insight on the factors that predict the probability of attaining population OS for these patients. The estimates from the different models varied widely at times. The CIs and CrIs for the log-ORs in the logistic models were generally very wide. While convergence was attained for these models, it was not possible to conclude that any of the covariates predicted the possibility of attaining the population OS. There were no differences between the Bayesian mixture and PTC models in terms of performance. The ML mixture model produced reasonable results to a large extent compared to the others and we would therefore recommend using the ML based mixture model when attempting to estimate π in this model.

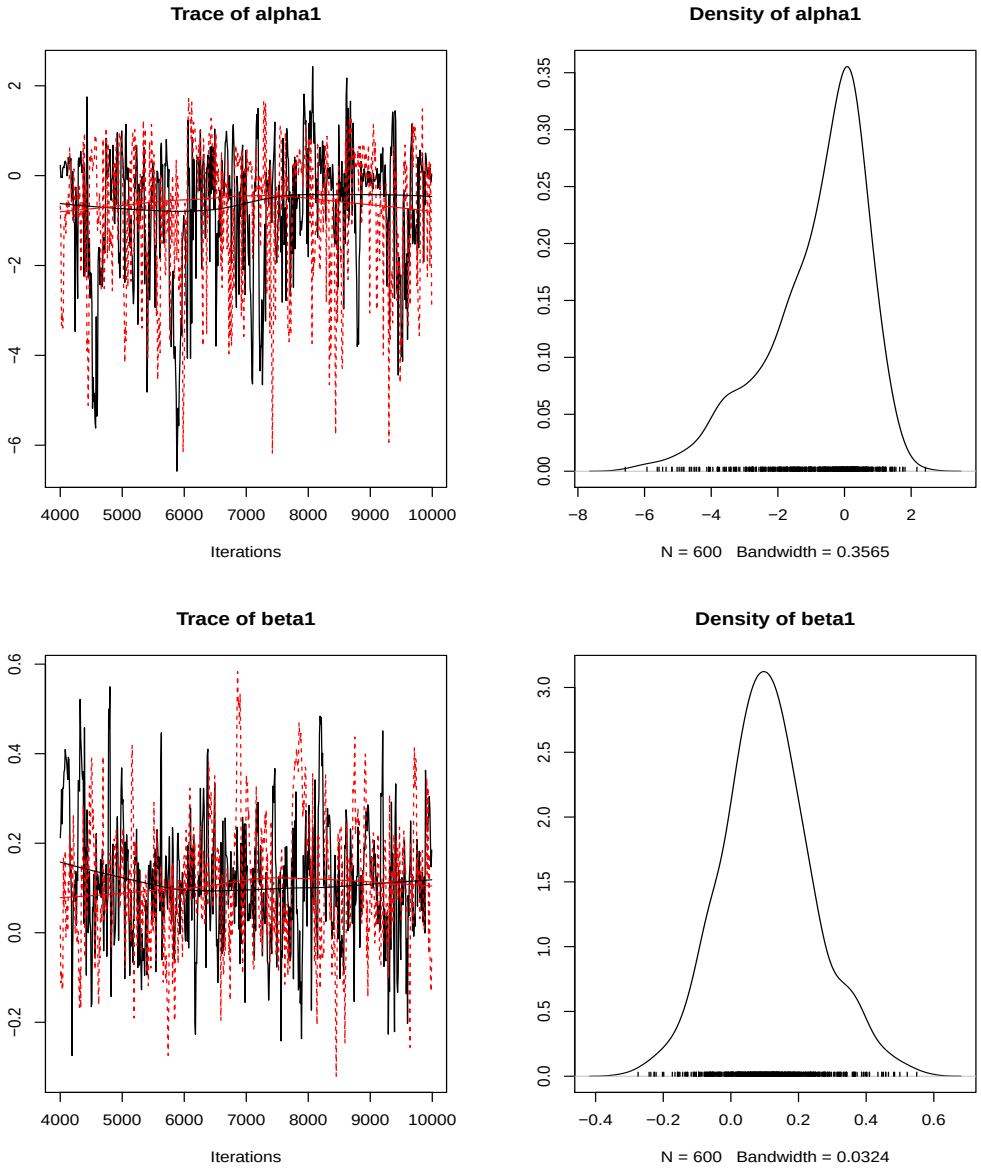


Figure 4.6: Trace and density plots from the Bayesian mixture model for OS fitted to the Myeloma dataset with paraprotein as a covariate on both π and λ. The model was fitted to data from 334 patients.

4.9 Missing data in cure rate models applied to Myeloma

The cure rate models fitted thus far included only individuals with complete data on each of the variables used in the analysis. To investigate whether estimates from mod-

els based on the CC analysis are similar to those based on the complete data likelihood as given in Equation 2.23, Bayesian models were used. The MI methods for combining outputs using Rubin's available in standard statistical software such as **mi estimate** in Stata do not presently allow for cure rate models. While it would be possible to fit models to each dataset and combine the outputs by hand, we used Bayesian methods to account for the missing data by treating them as unknown quantities for which we seek to elicit a posterior distribution on top of the usual model parameters we are interested in. We thus seek to specify a joint model for the observed and missing data including the model parameters in order to proceed with the estimation in MCMC. The Bayesian models had produced robust estimates when accounting for missing data in Section 3.4.5.

Paraprotein and beta2 were the two variables used in the cure rate models that had missing data as only one patient had a missing albumin measure while all 427 individuals had log-RD data. For the TTR, the Bayesian Weibull mixture model was used to estimate the proportion who would not relapse after treatment while the Bayesian Weibull population mixture model was used to model the proportion whose OS would return to that of disease-free individuals in the general population. The usual $N(0, 100)$ priors were used for the intercepts and coefficients of the effect of paraprotein and beta2 as independent covariates on π and the scale parameter λ for those not cured. On top of this, for each covariate with missing data, a $N(0, 100)$ prior distribution was specified for the mean of the missing observations, while a $Gamma(1, 1)$ prior was specified for the variance of the unobserved data [131]. Again, the missing data mechanism was assumed ignorable.

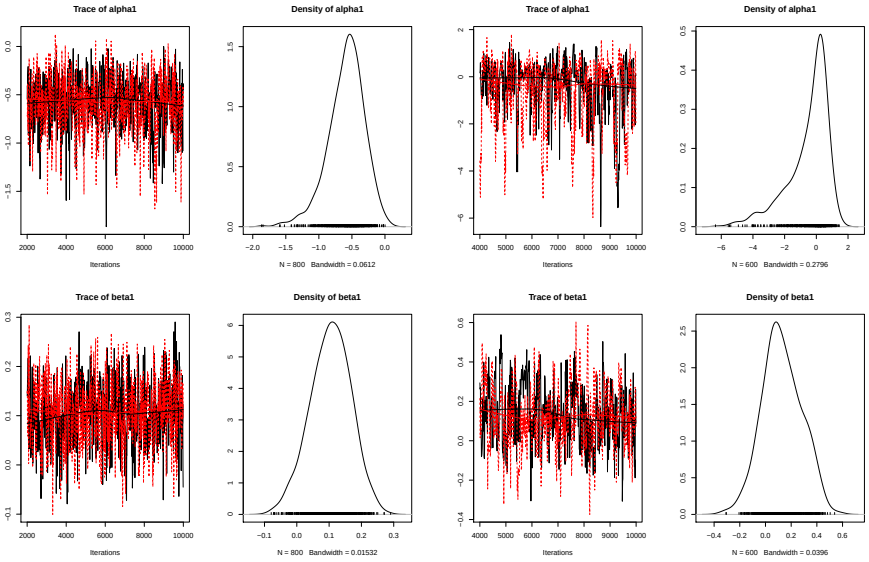
Estimates in the Bayesian mixture model to estimate the proportion who would never relapse following treatment were similar if not more precise than those from the CC analysis. For instance in the logistic model for π, the log-OR (95% CrI) for paraprotein in the CC Bayesian Weibull mixture model for TTR was -0.621 (-1.392, -0.135), while

it was -0.586 (-1.174, -0.143) after accounting for missing observations. With beta2 as the covariate on π, the estimated log-ORs (95% CrIs) were respectively -0.370 (-0.939, 0.043) and -0.386 (-0.900, 0.006), Table 4.5. Similarly, when estimating the proportion whose OS goes back to that of age-sex matched individuals in the general population, the log-ORs for paraprotein and beta2 were respectively -0.796 (-4.347, 1.231) and -0.805 (-3.413. 0.754) in the CC analysis, while in the models fitted to complete data, they were -0.517 (-4.034, 1.101) and -0.749 (-3.024, 0.983) respectively. The effect sizes were smaller in the models for π when the missing data was taken into account. On the other hand, the CrIs for the estimates in models fitted to the complete dataset were narrower implying more precision in the estimates. The models for TTR and OS yielded similar results to the models based on CCs, with a general improvement in the precision as evidenced by the narrower CrIs.

The SEs from the models accounting for missing data were bigger in the models for π than in the TTR or OS models with the biggest SEs seen in the Bayesian mixture model fitted to OS data. All models generally fitted well as evidenced by the trace and density plots for paraprotein as a covariate in the cured proportion and the TTR or OS models, Figure 4.7. The density plot for the log-OR for paraprotein in the Bayesian population mixture model spanned a whole range of values supporting the wide CrI for the estimate, confirming that it was hard to estimate π using the OS data in Myeloma as earlier highlighted.

4.10 Multivariable cure rate models in Myeloma

We investigated models with more than one covariate on the model for π as well as the TTR and OS. With TTR as the outcome, we used the smcure package in R and obtained an error showing that the model could not converge and the parameters could not be estimated even with just two covariates. A trace-back showed that in the logistic



A B

Figure 4.7: Trace and density plots of the coefficients for paraprotein in the Myeloma dataset on the cured proportion, alpha1 (top) with the effect on the TTR on the left (A) and on OS (B) on the right. Estimates of the effect of paraprotein (beta1) on the TTE (TTR or OS) are shown on the second row in the respective models.

model for TTR, the estimated parameters were too high such that $\exp(\alpha_0 + \alpha_1 z_i)$ was infinite which consequently led to a break in the code. We again tried this for the OS using the strsmix package in Stata. The results are shown in the Table 4.6 below. It is clear from these results that it was not possible to model the cured proportion while adjusting for other factors. The point estimates, SEs and CIs for the log-ORs in the cure model were very large. In contrast, the estimates from the OS model were plausible and similar to what had been obtained in the standard analysis. We thus conclude that we cannot model the OS using cure rate models that include more than one covariate in the cure model.

4.11 Comparing estimates in standard and mixture models in Myeloma

Having fitted univariable models with various covariates on both the cured proportion and the TTR, forest plots were used to check if the effect of log-RD changed having taken into account the proportion that may have been cured, Figures 4.8 and 4.9. Estimates from the standard Cox PH, Exponential and Weibull models were similar when modelling the TTR as seen previously. Further, the effect of log-RD was borderline significant in all models. When the possibility of a cure was taken into account however, there was no significant association of log-RD and the TTR as shown by the 95% CI, for the Cox PH mixture model and 95% CrIs for the Bayesian Weibull mixture and PTC models. The estimate of the log-HR was generally smaller in the cure rate models, suggesting that perhaps having taken into account those who are cured, the RD following treatment on its own was not the main predictor of TTR. Finally, the estimate from the PTC models was much smaller than that in the other two cure rate models based on the wider CrI.

The estimates from the standard models for OS were comparable when looking at log-RD. However, the estimates from the ML mixture, Weibull mixture and Weibull PTC models were more varied. Moreover, the CI for the ML Weibull mixture and the CrI for the Weibull PTC model were very wide. This emphasises the fact that it is difficult to identify the proportion whose OS is similar to that in the general population for the Myeloma data.

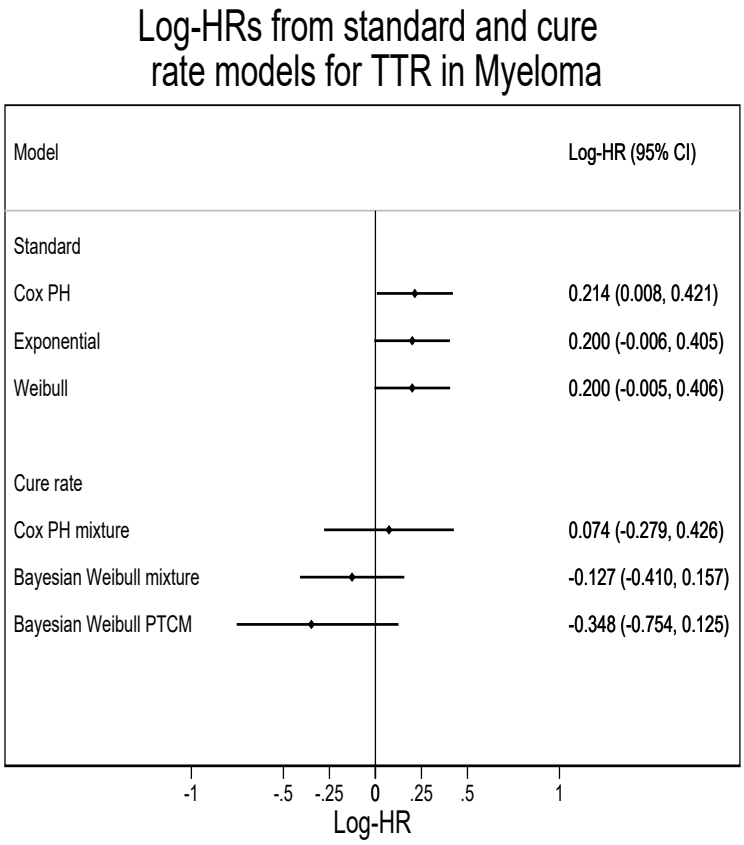


Figure 4.8: Estimates of the log-HR for log-RD in the TTR standard (top) and cure rate models (bottom). Dots represent point estimates of the log-HR while lines are the 95% CIs in the models fitted using ML or CrIs in the Bayesian models.

4.12 Cure rate models applied to the CLL dataset

Having fitted the cure rate models to the Myeloma dataset, the mixture and PTC models were applied to the CLL data to estimate the proportion that will never relapse following treatment and also that whose OS will return to that of the general population. As before, a logistic model for π and a Weibull model for TTR and OS was used.

The estimates for the proportion that will never relapse after treatment were 11 and 12% respectively in the mixture and PTC models without taking into account covariates

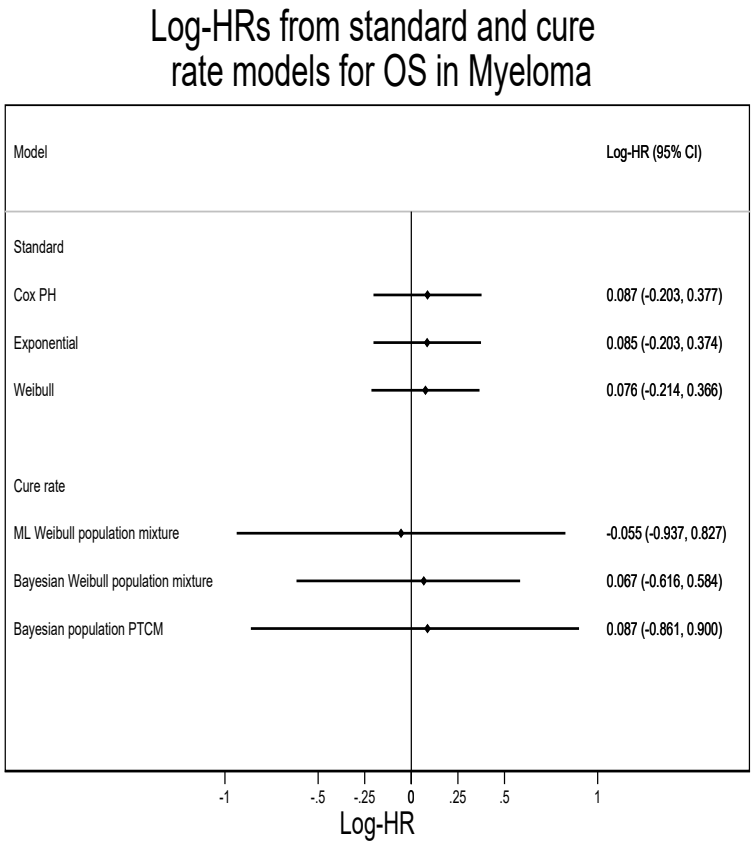
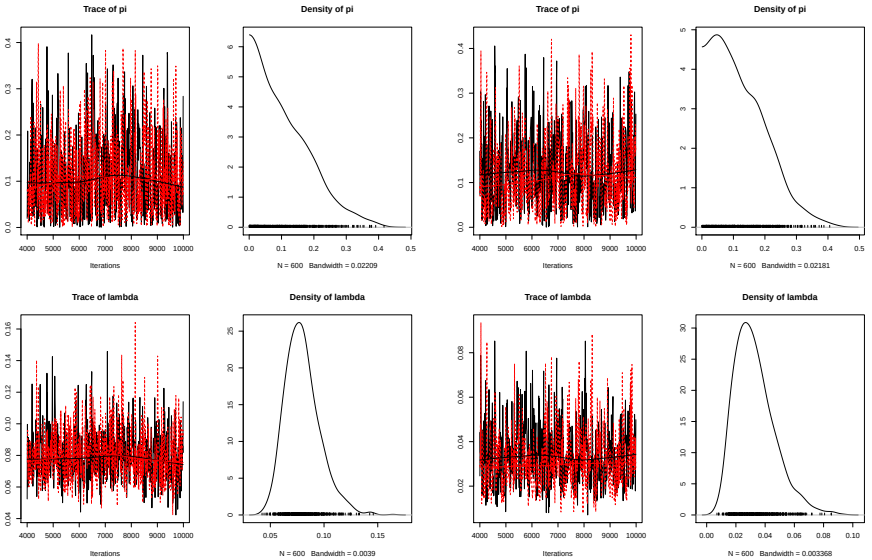


Figure 4.9: Estimates of the log-HR for log-RD in the OS standard (top) and cure rate models (bottom). Dots represent point estimates of the log-HR while lines are the 95% CIs in the models fitted using ML or CrIs in the Bayesian models.

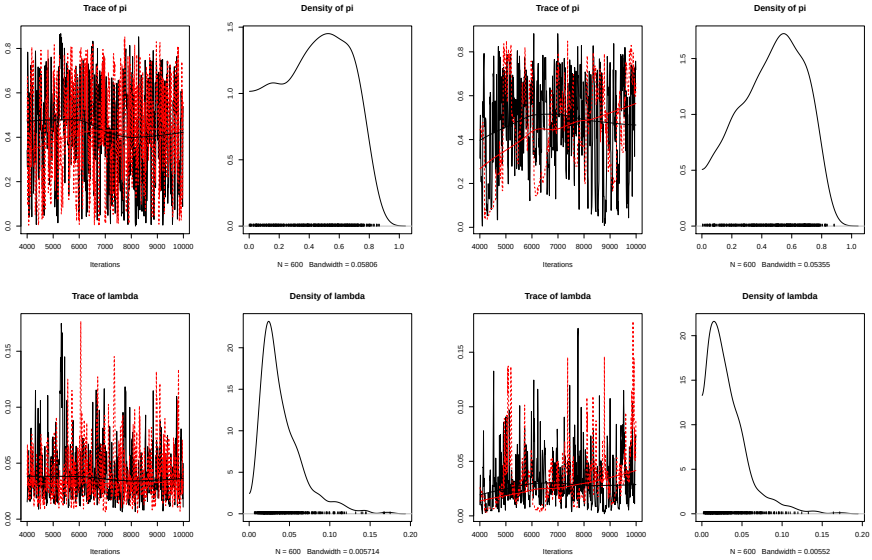
on π or λ in the Weibull model for TTR. However, both estimates had wide CIs, Table 4.7. Despite this, the conditional posterior distributions for each parameter converged to stationary distributions, Figure 4.10. The trace plots for both the Bayesian mixture and Bayesian PTC models showed that convergence was attained. However, the density plots suggested that for the TTR, the proportion cured and who would not relapse again in CLL could not be estimated precisely (top two density plots). Again worryingly, the density plots from the Bayesian mixture (left) and Bayesian PTC (right) spanned the whole range implying no statements about the identification of a proportion could be

made.

The population cure models depicted even higher estimates of π at 43 and 46% in the mixture and PTC models respectively. As before, these models also converged without problems despite the wide CrIs for the estimate of π, Figure 4.10. Despite the fact that almost 90% relapsed, the larger proportion who have OS similar to the age-sex matched general population suggests that many may ultimately die of other causes. The posterior distributions for π in these models were not centred around a particular value implying cure could not be identified precisely. This is more likely to be the case for a chronic disease such as CLL than for a more acute disease such as Myeloma. In the CLL dataset, the posterior density for π representing the proportion whose OS would return to that of disease-free individuals in the general population again spanned the whole range implying it would be hard to make statements about cure.



(A) Bayesian mixture (B) Bayesian PTC



(C) Bayesian population mixture (D) Bayesian population PTC

Figure 4.10: Trace and density plots for the cured proportion (π) and the scale parameter (λ) from the MCMC output in Bayesian Weibull mixture and PTC models (A and B) for TTR as well as Bayesian Weibull population mixture and PTC models (C and D) for OS fitted to the CLL dataset.

4.12.1 Covariates in cure rate models for OS in CLL

It was not possible to estimate the proportion who will not relapse following treatment in CLL using mixture or PTC model with covariates on π and λ. This was to be expected

as the empirical K-M curve for the TTR in CLL did not have an evident plateau from which π could be reliably identified.

In the population mixture and PTC models with covariates in CLL, log-RD, stage and sex were fitted as covariates on both π and λ, Table 4.8. A higher log-RD was indicative of a lower probability of attaining the population level OS in both models though this effect was not significant. With respect to OS, a higher log-RD was predictive of a shorter time to death although this effect was not significant in the PTC model. Stage and sex were not associated with π in both the mixture and PTC models. For the OS, those with a higher disease stage were more likely to have a shorter time to death, similar to the trend seen in the Weibull model for OS alone without assuming a proportion are cured. In contrast to the Weibull model for OS only, females were at a higher risk death among those not cured than males. However, this effect was not significant. Moreover, the CrIs in both models were very wide. In the CLL, the Bayesian PTC model for OS in general had estimates with narrower CrIs compared to the mixture model especially when estimating the log-ORs in the logistic model for π. However, both models depicted lack of fit through the trace and density plots in general.

4.13 Summary

The work thus far has focused on modelling the factors associated with TTR and OS using the Myeloma IX and CLL trial data, focusing on the role of RD and its logarithm, log-RD, on these outcomes, Chapter 3.

In this chapter, we examined the factors associated with being cured from Myeloma following treatment and also modelling the TTR and OS for those not cured in the mixture models, and the short-term effect of these covariates on the TTR and OS in the PTC models. In both malignancies, log-RD is an important predictor of TTR and OS and was thus included as a covariate in both the cure model and the TTR or OS model in the mixture and the PTC models. The K-M estimates for CLL for both TTR and OS showed curves without the requisite 'plateau' towards the end of follow-up from which the cured proportion could be estimated. The follow-up in CLL was around 5 years which may not have been sufficient to estimate the cured proportion. In Myeloma on the other hand, the K-M curve 'flattens out' after approximately 6 years for TTR. It was thus feasible to estimate the cured proportion for TTR in Myeloma. Mixture and PTC models have been used to model the time to progression in Myeloma where the possibility of a cured has been considered [22].

The models for estimating the proportion whose OS was similar to age-sex matched individuals from the general population showed that while it is feasible to get estimates of the cured proportion and model OS using the mixture and PTC models, the estimates had very wide CIs and CrIs and the estimate of π itself ranged from 2% to above 70%. The estimates from the PTC models were less reliable than those from mixture models (4.8) especially when estimating π in population mixture models which was in agreement with what has been shown [132].

For TTR, the Bayesian mixture model provided better estimates than the PH mixture and Bayesian PTC models when it came to estimating the proportion likely to have been

cured following treatment and therefore not bound to relapse as well as modelling the TTR. The PTC model did not perform better than the PH and Weibull mixture models in general when estimating the cured proportion and modelling the TTR. The formulation of the PTC model makes it more suitable for modelling the cured proportion where interest is on the number of residual malignant cells after treatment and because of that, it attaches biological meaning. Our focus in this thesis was on whether we could estimate the cured proportion in the Myeloma and CLL datasets. For this purpose, it would suffice to use the mixture models. It was not possible to adjust for other covariates in the cure rate models with TTR outcomes in the Myeloma dataset, while for CLL, it was not possible to model covariates at all. When applied to OS, the ML mixture model fitted in Stata performed better than the Bayesian population mixture and PTC models. However, the presence of a proportion whose OS returned to that of disease-free individuals in the general population could not be ascertained. Further, modelling more than one covariate on the cure proportion did not produce meaningful results even when using the ML population mixture model.

Finally, to account for missing data, we modelled the missing data by specifying prior distributions for the covariates with missing data in a Bayesian setting. The models accounting for missing data gave better estimates and were more informative regarding the size of the effect of paraprotein and beta on the cured proportion and the TTR and OS.

Having estimated the cured proportion using the mixture and PTC models, it was evident that the log-RD is not the only important predictor of the cured proportion and TTR or OS. Another important tumour characteristic to take into account when estimating π is the growth rate of the RD after treatment. In both data sets, this was not recorded, however for Myeloma, data was available on covariates that reflect the tumour growth characteristics. It is thus feasible to model the effect of the tumour growth, as a latent variable on the cured proportion and the TTR or OS using the structural equation

modelling technique which is discussed in the next chapter.

	PH mixture (TTR)	**Mixture model (TTR)**	**PTC model (TTR)**
Parameter	Estimate (95% CI)	Estimate (95% CrI)	Estimate (95% CrI)
Cure model			
Intercept	-2.201 (-2.797, -1.605)	-1.968 (-2.546, -1.431)	-2.131 (-2.992, -1.512)
Log-RD	-0.804 (-1.455, -0.154)	-0.739 (-1.228, -0.206)	-0.791 (-1.457, -0.152)
TTR model			
Intercept	-	-1.023 (-1.275, -0.782)	-1.886 (-2.301, -1.529)
Log-RD	0.074 (-0.279, 0.426)	-0.127 (-0.410, 0.157)	-0.348 (-0.754, 0.125)
Shape (γ)	-	1.301 (1.164, 1.437)	1.423 (1.264, 1.575)
DIC	-	12,580	12,577
Cure model			
Intercept	-1.845 (-3.017, -0.673)	-1.702 (-2.394, -1.257)	-2.032 (-3.174, -1.389)
Paraprotein	-0.668 (-2.560, 1.223)	-0.621 (-1.392, -0.135)	-0.730 (-1.749, -0.036)
TTR model			
Intercept	-	-1.041 (-1.229, -0.861)	-1.908 (-2.400, -1.586)
Paraprotein	0.203 (0.035, 0.372)	0.111 (-0.031, 0.227)	-0.033 (-0.298, 0.248)
Shape (γ)	-	1.296 (1.134, 1.458)	1.392 (1.212, 1.561)
DIC	-	9,844	9,843
Cure model			
Intercept	-1.806 (-2.275, -1.337)	-1.652 (-2.208, -1.249)	-1.885 (-3.082, -1.302)
Beta2	-0.528 (-1.063, 0.006)	-0.370 (-0.939, 0.043)	-0.417 (-1.123, 0.048)
TTR model			
Intercept	-	-1.033 (-1.225, -0.846)	-1.855 (-2.302, -1.546)
Beta2	0.087 (-0.127, 0.302)	0.138 (0.006, 0.262)	0.098 (-0.168, 0.293)
Shape (γ)	-	1.307 (1.144, 1.461)	1.430 (1.216, 1.629)
DIC	-	9,142	9,138
Cure model			
Intercept	-1.818 (-3.882, 0.245)	-1.513 (-1.984, -1.172)	-1.855 (-3.756, -1.273)
Albumin	0.768 (-1.098, 2.634)	0.172 (-0.221, 0.569)	0.057 (-1.004, 0.642)
TTR model			
Intercept	-	-0.967 (-1.147, -0.800)	-1.771 (-2.426, -1.456)
Albumin	-0.134 (-0.382, 0.112)	-0.137 (-0.316, 0.042)	-0.191 (-0.615, 0.183)
Shape (γ)	-	1.277 (1.147, 1.418)	1.367 (1.163, 1.558)
DIC	-	12,551	12,548

Table 4.2: Log-RD ($n = 427$), paraprotein ($n = 334$), beta2 ($n = 310$) and albumin ($n = 426$) as independent covariates in mixture and PTC models for TTR in Myeloma. In the cure models, estimates of the intercept and log-ORs from all models are presented, while estimates from the TTR models include the log-HRs in all models, and log-HRs as well as the intercept and shape parameter (γ) in the Weibull models.

	ML mixture	**Bayesian mixture model**	**Bayesian PTC**
Parameter	Estimate π (95% CI)	Estimate π (95% CrI)	Estimate of π (95% CrI)
Cure model			
π	0.308 (0.056, 0.770)	0.278 (0.025, 0.508)	0.290 (0.033, 0.503)
OS model			
λ	0.082 (0.043, 0.156)	0.084 (0.052, 0.126)	0.049 (0.017, 0.090)
γ	1.390 (1.103, 1.753)	1.373 (1.136, 1.645)	1.439 (1.194, 1.711)
DIC		12,300	12,300

Table 4.3: Estimates of π and the scale (λ) and shape (γ) parameters from Weibull population mixture and PTC models for OS in the Myeloma dataset fitted using ML and MCMC (Bayesian)

	ML mixture (OS)	**Mixture model (OS)**	**PTC model (OS)**
Parameter	Estimate (95% CI)	Estimate (95% CrI)	Estimate (95% CrI)
Cure model			
Intercept	-0.893 (-2.270, 0.914)	-2.465 (-7.123, -0.109)	-1.559 (-4.574, -0.116)
Log-RD	-0.313 (-1.721, 1.095)	-0.471 (-1.788, 4.187)	0.138 (-1.173, 1.978)
OS model			
Intercept	-2.498 (-3.169, -1.827)	-2.618 (-3.159, -2.071)	-3.327 (-4.366, -2.474)
Log-RD	-0.055 (-0.937, 0.827)	0.067 (-0.616, 0.584)	0.087 (-0.861, 0.900)
Shape (γ)	1.407 (1.114, 1.775)	1.347 (1.115, 1.617)	1.430 (1.181, 1.706)
DIC	-	12,301	12,301
Cure model (π)	Log-OR (95% CI)	Log-OR (95% CrI)	Log-OR (95% CrI)
Intercept	-1.215 (-4.487, 2.056)	-3.201 (-7.677, -0.144)	-1.961 (-5.757, -0.151)
Paraprotein	0.285 (-0.488, 1.058)	-0.796 (-4.347, 1.231)	-0.066 (-1.782, 1.063)
OS model			
Intercept	-2.821 (-3.552, -2.090)	-2.906 (-3.452, 2.376)	-3.733 (-4.858, -2.767)
Paraprotein	0.269 (-0.018, 0.556)	0.110 (-0.151, 0.401)	0.145 (-0.457, 0.649)
Shape (γ)	1.470 (1.131, 1.912)	1.417 (1.175, 1.710)	1.529 (1.258, 1.855)
DIC	-	9,604	9,604
Cure model			
Intercept	-0.449 (-1.658, 0.760)	-2.115 (-6.161, 0.005)	-1.287 (-3.451, 0.066)
Beta2	-0.269 (-0.769, 0.231)	-0.805 (-3.413. 0.754)	-0.593 (-1.970, 0.408)
OS model			
Intercept	-2.606 (-3.204, -2.008)	-2.769 (-3.325, -2.190)	-3.401 (-4.359, -2.564)
Beta2	0.277 (0.029, 0.524)	0.236 (-0.043, 0.542)	0.120 (-0.332, 0.602)
Shape (γ)	1.565 (1.245, 1.967)	1.434 (1.181, 1.753)	1.546 (1.236, 1.879)
DIC	-	8,918	8,917
Cure model			
Intercept	-1.288 (-3.201, 0.624)	-2.330 (-6.070, -0.355)	-1.501 (-3.545, -0.135)
Albumin	-0.655 (-1.560, 0.251)	-0.673 (-2.188 1.294)	-0.306 (-1.327, 1.290)
OS model			
Intercept	-2.638 (-3.095, -2.182)	-2.704 (-3.070, -2.320)	-3.376 (-4.119, -2.611)
Albumin	-0.494 (-0.758, -0.231)	-0.360 (-0.708, 0.071)	-0.435 (-0.963, 0.447)
Shape (γ)	1.415 (1.166, 1.717)	1.371 (1.160, 1.625)	1.463 (1.218, 1.764)
DIC	-	12,265	12,266

Table 4.4: Log-RD ($n = 427$), paraprotein ($n = 334$), beta2 ($n = 310$), albumin ($n = 426$) as independent covariates in mixture and PTC models for OS in Myeloma. In the cure models, estimates of the intercept and log-ORs from all models are presented, while estimates from the OS models include the log-HRs in all models, and log-HRs as well as the intercept and shape parameter (γ) in the Weibull models.

	Bayesian mixture model (TTR)		**Bayesian mixture model (OS)**	
Cure model	Estimate (95% CrI)	SE	Estimate (95% CrI)	SE
Cure model				
Intercept	-1.559 (-2.013, -1.216)	0.005	-2.807 (-7.161, -0.356)	0.051
Paraprotein	-0.586 (-1.174, -0.143)	0.007	-0.517 (-4.034, 1.101)	0.038
TTE model				
Intercept	-0.967 (-1.142, -0.809)	0.002	-2.703 (-3.102, -2.255)	0.006
Paraprotein	0.107 (-0.037, 0.223)	0.002	0.128 (-0.167, 0.422)	0.004
Shape (γ)	1.292 (1.161, 1.429)	0.002	1.333 (1.134, 1.590)	0.003
DIC	13,518		13,247	
Cure model				
Intercept	-1.505 (-1.927, -1.169)	0.005	-2.524 (-2.725, 0.261)	0.053
Beta2	-0.386 (-0.900, 0.006)	0.006	-0.749 (-3.024, 0.983)	0.027
TTE model				
Intercept	-0.967 (-1.132, -0.799)	0.002	-2.725 (-3.176, -2.204)	0.007
Beta2	0.133 (0.008, 0.243)	0.002	0.261 (0.021, 0.513)	0.004
Shape (γ)	1.290 (1.159, 1.429)	0.002	1.372 (1.153, 1.642)	0.004
DIC	13,454		13,163	

Table 4.5: Estimates of log-OR and log-HRs for paraprotein and beta2 as covariates in Bayesian mixture models for TTR and OS in Myeloma with imputation models to account for missing data. In the cure models, estimates of the intercept and log-ORs are presented, while estimates from the TTE models include log-HRs as well as the intercept and shape parameter (γ) in the Weibull models.

	ML mixture model (OS)	
Parameter	Estimate (SE)	(95% CI)
Cure model		
Intercept	-1613.1 (98987.4)	(-195624.7, 192398.6)
Log-RD	1179.2 (66769.7)	(-129687.1, 132045.5)
Paraprotein	-709.8 (42903.9)	(-84799.8, 83380.3)
Beta2	154.8 (7314.8)	(-14182.0, 14491.7)
Albumin	-1307.1 (78535.0)	(-155232.8, 152618.6)
OS model		
Intercept	-2.981 (0.304)	(-3.577, -2.384)
Log-RD	0.401 (0.259)	(-0.106, 0.908)
Paraprotein	0.093 (0.121)	(-0.144, 0.329)
Beta2	0.295 (0.090)	(0.119, 0.471)
Albumin	-0.382 (0.146)	(-0.669, -0.095)
Shape (γ)	Not shown	Not shown

Table 4.6: Log-RD, paraprotein, beta2 and albumin as covariates in a multivariable ML mixture model for OS in Myeloma fitted to data from $n = 243$ patients. In the cure model, log-OR estimates as coefficients of each covariate, the SE and 95% CI are reported, while estimates in the OS model represent log-HRs, their SE and 95% CI.

	Mixture model	**PTC model**
Parameter	Estimate (95% CrI)	Estimate (95% CrI)
Cure model		
π	0.111 (0.004, 0.322)	0.122 (0.007, 0.317)
TTR model		
λ	0.080 (0.055, 0.118)	0.033 (0.013, 0.066)
γ	1.579 (1.324, 1.847)	1.647 (1.402, 1.931)
DIC	9,235	9,237
Cure model		
π	0.426 (0.027, 0.798)	0.464 (0.053, 0.796)
OS model		
λ	0.040 (0.010, 0.115)	0.032 (0.005, 0.102)
γ	1.657 (1.120, 2.402)	1.672 (1.120, 2.341)
DIC	8,910	8,911

Table 4.7: Bayesian mixture and PTC models for π and TTR and OS with a Weibull model in the CLL dataset. Estimates of π, the scale (λ) and the shape (γ) are reported for each model.

	Bayesian mixture (OS)	**Bayesian PTC (OS)**
Parameter	Estimate (95% CrI)	Estimate (95% CrI)
Cure model		
Intercept	-1.180 (-3.700, 0.990)	-1.160 (-4.100, 1.110)
Log-RD	-0.430 (-1.890, 0.640)	-0.380 (-1.700, 0.660)
OS model		
Intercept	-3.990 (-5.160, -2.810)	-4.670 (-6.290, -2.950)
Log-RD	0.550 (0.090, 0.980)	0.480 (-0.160, 1.090)
Shape (γ)	1.690 (1.140, 2.470)	1.750 (1.170, 2.610)
Cure model		
Intercept	-0.710(-3.560, 1.560)	-0.730 (-3.420, 1.590)
Stage II	-0.540 (-3.640, 2.240)	-0.520 (-3.660, 2.180)
OS model		
Intercept	-3.840 (-5.370, -2.280)	-4.480 (-6.320, -2.510)
Stage II	0.240 (-1.330, 1.560)	0.180 (-1.940, 2.030)
Shape (γ)	1.770 (1.150, 2.640)	1.830 (1.180, 2.820)
Cure model		
Intercept	-0.590 (-3.510, 1.410)	-0.660 (-3.110, 1.150)
Sex	1.130 (-2.190, 4.110)	0.930 (-2.360, 3.910)
OS model		
Intercept	-3.720 (-5.410, -2.240)	-4.310 (-6.080, -2.820)
Sex	0.330 (-1.570, 2.500)	0.490 (-2.000, 3.280)
Shape (γ)	1.840 (1.140, 2.890)	1.880 (1.210, 2.910)

Table 4.8: Log-RD, stage and sex as covariates in univariable mixture and PTC models for those whose OS returns to that of disease-free individuals in the CLL dataset. In the cure models, estimates of the intercept and log-ORs from both models are presented, while estimates from the OS models include the log-HRs as well as the intercept and shape parameter (γ) in the Weibull models.

Chapter 5

Using structural equation modelling to evaluate the effect of a latent covariate on TTE outcomes

5.1 Introduction

This chapter discusses methods for modelling how latent tumour characteristics such as tumour re-growth rate and RD disease after treatment influence TTE outcomes and demonstrate how these models are fitted using available software.

The use of latent variables in modelling survival outcomes has not been widely discussed in the literature mainly because SEM approaches have not traditionally been used in medical research. In recent times, there has been a growing interest in using SEMs to model latent variables in a number of disease areas. For instance, in liver cancer, SEMs were used to model drug responses using observed biomarkers [133]. Other applications in lung cancer have involved examining whether the observed risk factors act independently or through complex mechanisms likely to be identified in a SEM framework [134]. This feature might be useful in our case where we have biomarkers which we believe are related to tumour growth which is itself an important predictor of

TTR and/OS. Rather than modelling the effect of each biomarker independently on the outcomes, we can explore within a SEM framework, whether these biomarkers together represent some latent construct which is itself a predictor of the survival outcomes. SEM extensions that enable the modelling of latent constructs through both binary and continuous observed variables on survival data in both parametric and semi-parametric TTE models have been discussed [135]. Again with survival outcomes, SEMs have been used for their greater explanatory power over independent covariates in data with small sample sizes but having high dimensionality [136]. SEMs that allow for a proportion who might be cured have been explored by considering the cured fraction as one of latent classes. However, these methods assumed a discrete time survival model [137]. The discrete time survival model bins the observed TTE into discrete chunks during which the event of interest could occur. In this thesis, we are interested in continuous survival times and a latent variable related to the TTE via continuous observed variables. We also aim to further extend the methods to model the effect of the continuous latent variable in cure rate models. We briefly discuss the general SEM framework.

Various methods have been proposed for modelling latent variables, some of which include data augmentation techniques such as the EM algorithm which have been discussed [74]. For this work, we have demonstrated that the cured proportion is itself an unobserved measure which can be modelled using cure rate models [18]. To incorporate latent covariates, SEMs may be used to elicit them from a set of observed variables that are known to be related to the latent measures. In both the Myeloma and CLL trials, there were no measurements of tumour growth. However, using bio-markers that are known to predict some characteristics of tumour growth, we can use SEMs to model how the tumour growth is associated with the outcomes TTR and OS and also the probability of being cured.

The SEM approach forms part of a broader range of models whose aim is to validate relationships based on observed data. This methodology extends generalised linear

modelling approaches such as analysis of variance and linear regression. Briefly, SEMs model latent variables by assuming a multivariate distribution for the observed data [53]. The relationship between unobserved latent variables and the manifest variables (observed data) is modelled by examining the covariance structure of the proposed model and comparing it to that obtained from the observed data through the covariance matrix. As a result, most of the methods for fitting SEMs do not require individual observations of the latent variable when using ML based techniques.

SEMs themselves are part of a wider group of models called latent variable models that fall into three main categories: 1) classical latent variable models also called common factor models where continuous observed variables depict an underlying continuous latent variable, 2) item response models where categorical observed variables relate to a latent continuous variable, and 3) latent class models where observed categorical variables relate to other latent categorical variables. This work is focused on classical latent variable models and how they can be extended to model TTE outcomes.

The general SEM has two components: 1) the *measurement model*, which describes how latent or unobserved variables are related to observed or manifest random variables, where the observed random variables are said to be manifestations of the unobserved latent variable and 2) the *structural model* which describes the relationships between latent variables [138]. In our case, the measurement model might relate the biomarkers albumin, beta2, and paraprotein to the latent variable reflecting tumour growth, whilst the structural model represents the TTE or cure rate model.

5.1.1 The measurement model

In the general case where there are several latent variables, the measurement model can be presented as follows. Let $\boldsymbol{y} = (y_1, \ldots, y_k)^T$ be a $k \times 1$ vector of observed continuous random variables which are manifestations of the continuous latent variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)^T$. Assume also that the observed manifest variables relate to the latent

variables through the measurement equation

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\Upsilon}\boldsymbol{\omega} + \boldsymbol{\epsilon}. \tag{5.1}$$

Here, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi}_{\boldsymbol{\epsilon}})$ where $\boldsymbol{\Psi}_{\boldsymbol{\epsilon}}$ is a diagonal matrix of the errors in the measurement of the variables $\boldsymbol{y}$, $\boldsymbol{\Upsilon}$ is a $k \times q$ matrix of coefficients or factor loadings to be estimated which describe the relationship between the latent variables and the observed variables while $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ is a $k \times 1$ vector of intercepts in the regression equations in (5.1). The latent scores themselves are assumed to be $N(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is the covariance matrix of the latent variables $\boldsymbol{\omega}$.

5.1.2 Measurement model with binary observed variables

In some instances, we may observe binary random variables that are related to some unknown latent variable $\boldsymbol{\omega}$ as before. In SEM, we assume the observed variables $\boldsymbol{y}$ represent some latent continuous responses $\boldsymbol{y}^*$ underlying the dichotomous response with

$$\boldsymbol{y} = \begin{cases} 1 & \text{if } \boldsymbol{y}^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

Then we can specify the model for $\boldsymbol{y}^*$ as in the case of continuous observed SEMs (5.1)

$$\boldsymbol{y}^* = \boldsymbol{\mu} + \boldsymbol{\Upsilon}\boldsymbol{\omega} + \boldsymbol{\epsilon} \tag{5.3}$$

The parameters for this model can be estimated as before by specifying a link function between the observed data and the latent variables.

5.1.3 The structural model

In SEM, interest might also be on how latent variables relate to each other. Structural models are used to represent these relationships. The structural model can be expressed as

$$\boldsymbol{\eta} = \boldsymbol{\nu} + \boldsymbol{B}\boldsymbol{\omega} + \boldsymbol{\zeta} \tag{5.4}$$

where $\boldsymbol{\eta}$ are latent outcome variables and $\boldsymbol{\omega} \sim N(\mathbf{0}, \boldsymbol{\Phi})$ as before, $\boldsymbol{\nu}$ are intercepts, $\boldsymbol{B}$ is an $r \times q$ non-singular matrix, while $\boldsymbol{\zeta}$ is a matrix of the errors in the structural model. For our purposes, interest is in a TTE outcome in the structural model (5.4) above so that we have

$$\boldsymbol{G(t)} = \boldsymbol{\nu} + \boldsymbol{B}\boldsymbol{\omega} \tag{5.5}$$

where $G(t)$ could be some function of the TTE such as the log-hazard function or the logistic model for π. The error term is therefore not required.

5.2 Estimating parameters in SEMs

There are two main approaches to estimating the parameters in SEMs: 1) the frequentist approach where the latent variables are treated as random and the parameters as fixed or in other instances, where both the latent variables and the parameters are treated as fixed, mainly to ease computation, and 2) a Bayesian approach where both the latent variables and the parameters are treated as random variables and inference is based on drawing samples from the posterior distribution in an MCMC.

5.2.1 Maximum likelihood estimation for SEMs

To fit the measurement model using maximum likelihood, consider the full likelihood based on data from n individuals given by

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\omega}, \boldsymbol{\Psi_\epsilon}; \boldsymbol{y}) &= \textstyle\prod_{i=1}^{n} f(\boldsymbol{y_i}|\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\omega_i}, \boldsymbol{\Psi_\epsilon}) \\ &= \textstyle\prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} |\boldsymbol{\Psi_\epsilon}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\epsilon_i}^T \boldsymbol{\Psi_\epsilon}^{-1} \boldsymbol{\epsilon_i}\right\} \end{aligned} \tag{5.6}$$

where $\boldsymbol{\epsilon_i} = \boldsymbol{y_i} - \boldsymbol{\mu} - \boldsymbol{\Upsilon}\boldsymbol{\omega_i}$ from (5.1). For this model, we can integrate out the latent scores $\boldsymbol{\omega_i}$ to have a likelihood based only on the observed data as

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}; \boldsymbol{y}) &= \textstyle\int \prod_{i=1}^{n} f(\boldsymbol{y_i}|\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\omega_i}, \boldsymbol{\Psi_\epsilon}) f(\boldsymbol{\omega_i}, \boldsymbol{\Phi}) \partial \boldsymbol{\omega_i} \\ &= \textstyle\prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y_i} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu})\right\} \end{aligned} \tag{5.7}$$

where $f(\boldsymbol{\omega_i}, \boldsymbol{\Phi})$ is the distribution of the latent variables and $\boldsymbol{\Sigma} = \boldsymbol{\Upsilon}\boldsymbol{\Phi}\boldsymbol{\Upsilon}^{\boldsymbol{T}} + \boldsymbol{\Psi_\epsilon}$. To estimate the parameters, one would have to maximise the likelihood or minimise the LL

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{Y}) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu}) \tag{5.8}$$

which can also be expressed as

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = n \log |\boldsymbol{\Sigma}| \; + \; n \; tr[\boldsymbol{\Sigma^{-1}}(\boldsymbol{S} \; + \; (\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})^{\boldsymbol{T}})] \tag{5.9}$$

where

$$S = \sum_{i=1}^{n} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})^T (\boldsymbol{y_i} - \bar{\boldsymbol{y}}) \tag{5.10}$$

is the observed sample covariance matrix. From this, the ML estimates for the parameters can be found by evaluating the score functions. For the mean $\boldsymbol{\mu}$, it is straightforward to evaluate

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y})}{\partial \boldsymbol{\mu}} = \mathbf{0}$$

due to the assumed normality of the distributions which results in the estimate $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$.

Estimating the variance components

To get estimates for $\boldsymbol{\Sigma}$, we need to minimize $l(\hat{\boldsymbol{\mu}},\ \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ [139]. As we only want elements that involve $\boldsymbol{\Sigma}$, the LL therefore becomes

$$l(\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}; \boldsymbol{y}) = n \log |\boldsymbol{\Sigma}| \ + \ n\ tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) \tag{5.11}$$

The estimation of the coefficients, correlations and variances is not usually possible unless some parameters of the model are constrained [55]. It has been shown that if $\boldsymbol{y}$ has a multivariate normal distribution, then S follows a Wishart distribution with n degrees of freedom [140]. Moreover, since in the model definitions $\boldsymbol{\Upsilon}$and $\boldsymbol{\Phi}$ are not independent, that is, the variance of $\boldsymbol{\omega}$ also depends on the intercepts or factor loadings, we need to impose restrictions on $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Phi}$ to ensure that we obtain unique estimates. Two well-known approaches involve fixing one or more of the elements in $\boldsymbol{\Upsilon}$ at known constants or fixing the variance of the latent variable(s), $\boldsymbol{\Phi}$. The latent variable is often assumed to have zero mean with the variance fixed at 1 to ensure identifiability or the coefficients. Numerical procedures have been proposed for finding the maximum likelihood estimates implemented in the standard packages already discussed are available [141].

5.2.2 Using Bayesian methods to fit SEMs

An attractive feature of the Bayesian approach is that it allows for the incorporation of additional information through prior distributions. This is done for each of the model unknowns, including the latent variables and the parameters from the measurement

and structural models. Inference is based on both the prior distribution and the observed data through the likelihood function. The prior allows external information about structural relationships, which may be available from previous studies or through expert knowledge, to be incorporated. Even if there is no prior knowledge, vague priors are used and the resulting estimates will still be comparable to those based on frequentist methods for large enough sample sizes [142].

With MCMC, it is possible to fit complex Bayesian models allowing for non-linearity, interactions and missing data, provided that full conditional distributions have standard forms. It is also straightforward to fit models with mixed categorical, count, and continuous observed variables. For this project, we implement SEMs that have TTE outcomes in the structural part. A problem with the MCMC approach is that being computationally intensive, it takes a long time to fit models and this can be a problem because SEMs are generally complicated models. Various computational tricks such as exploring the geometry of the target distribution before constructing the algorithm or using scalable algorithms that break up the problem into manageable pieces among several other techniques, can be used to obtain estimates within an acceptable level of error [143]. It was not necessary to use such tricks when fitting models in this work.

A final important feature is that we are able to obtain from the joint posterior distribution, a sample from the conditional distribution of the latent variables themselves. These samples can then be used to obtain important insights into structural relationships, which may not be apparent from just looking at the parameter estimates.

5.2.3 Bayesian estimation of parameters in measurement models

To illustrate how SEMs are fitted, consider the measurement model as defined in (5.1) and assume we observe data from n individuals so that

$$\boldsymbol{y_i} = \boldsymbol{\mu} + \boldsymbol{\Upsilon}\boldsymbol{\omega_i} + \boldsymbol{\epsilon_i}. \tag{5.12}$$

where $i = 1, \ldots, n$, and $\boldsymbol{\mu}$, $\boldsymbol{\Upsilon}$ and $\boldsymbol{\epsilon_i}$ are defined as before. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Phi}, \boldsymbol{\Psi_\epsilon})$ be a vector of the unknown parameters in the model to be estimated. Further, let $\boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$ be a matrix of the observed random variables related to the matrix of the latent scores $\boldsymbol{\Omega} = (\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega_n})$. By treating the the latent scores $\boldsymbol{\Omega}$ as hypothetical missing data, the complete data set is $(\boldsymbol{\Omega}, \boldsymbol{Y})$. It is then possible to generate samples of $(\boldsymbol{\Omega}, \boldsymbol{Y})$ from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\Omega}|\boldsymbol{Y})$ by considering the full conditional distributions of the latent scores and model parameters using such MCMC techniques as the Gibbs sampling algorithm as detailed in Section 2.9.3.

Full conditional distribution for the latent scores $\boldsymbol{\Omega}$

To derive the conditional distribution $p(\boldsymbol{\Omega}|\boldsymbol{\Psi_\epsilon}, \boldsymbol{\Upsilon}, \boldsymbol{\Phi}, \boldsymbol{Y}) = p(\boldsymbol{\Omega}|\boldsymbol{\theta}, \boldsymbol{Y})$, note that the $\boldsymbol{\omega_i}$ are mutually independent for $i = 1, \ldots, n$ and that observed data $\boldsymbol{y_i}$ are also independent given $(\boldsymbol{\omega_i}, \boldsymbol{\theta})$. The conditional posterior distribution of $\boldsymbol{\Omega}$ given $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ can then be expressed as

$$p(\boldsymbol{\Omega}|\boldsymbol{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\omega_i}|\boldsymbol{y_i}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(\boldsymbol{y_i}|\boldsymbol{\omega_i}, \boldsymbol{\theta}) p(\boldsymbol{\omega_i}|\boldsymbol{\theta}). \tag{5.13}$$

Since from the model definition in (5.1), $\boldsymbol{\omega_i}|\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{\Phi})$ and $\boldsymbol{y_i}|\boldsymbol{\omega_i}, \boldsymbol{\theta} \sim N(\boldsymbol{\mu} + \boldsymbol{\Upsilon}\boldsymbol{\omega_i}, \boldsymbol{\Psi_\epsilon})$, we can derive the full conditional distribution of $\boldsymbol{\omega_i}$ given $(\boldsymbol{y}, \boldsymbol{\theta})$ as

$$\begin{aligned} p(\boldsymbol{\Omega}|\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}, \boldsymbol{\Phi}) \propto \exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\boldsymbol{y_i} - (\boldsymbol{\mu} + \Upsilon\boldsymbol{\omega_i}))^T \Psi_\epsilon^{-1}(\boldsymbol{y_i} - (\boldsymbol{\mu} + \Upsilon\boldsymbol{\omega_i}))\right\} \\ \times \exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n} \boldsymbol{\omega_i}^T \Phi^{-1} \boldsymbol{\omega_i}\right\}. \end{aligned} \tag{5.14}$$

Conditional distributions for the elements in θ

Having determined the conditional distribution of $\boldsymbol{\Omega}$, we can now work out the conditional distributions of the remaining parameters in $\boldsymbol{\theta}$ given the complete data $(\boldsymbol{\Omega}, \boldsymbol{Y})$.

For the prior for $\boldsymbol{\theta}$, we have

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}) p(\boldsymbol{\Phi}) \tag{5.15}$$

and since from the measurement equation in (5.1) the observed data $\boldsymbol{Y}$ depends only on $\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}$ and by definition, the distribution of $\boldsymbol{\Omega}$ only depends on $\boldsymbol{\Phi}$. The joint posterior can then be factored as follows.

$$p(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}, \boldsymbol{\Phi} | \boldsymbol{Y}, \boldsymbol{\Omega}) = \{p(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}) p(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon})\} \times p(\boldsymbol{\Omega} | \boldsymbol{\Phi}) p(\boldsymbol{\Phi}) \tag{5.16}$$

This is convenient since we can now look at the conditional densities separately. Having factored the joint posterior distribution, we can separately work out the marginal conditional densities

$$p(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon} | \boldsymbol{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}) p(\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon})$$

and

$$p(\boldsymbol{\Phi} | \boldsymbol{Y}, \boldsymbol{\Omega}) \propto p(\boldsymbol{\Omega} | \boldsymbol{\Phi}) p(\boldsymbol{\Phi})$$

which makes it possible to specify prior distributions for $\boldsymbol{\mu}, \boldsymbol{\Upsilon}, \boldsymbol{\Psi_\epsilon}$ and $\boldsymbol{\Phi}$. For instance we can have for $\boldsymbol{\mu}$

$$\boldsymbol{\mu} \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$$

while for $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Psi_\epsilon}$, consider the j-th row and diagonal elements $\boldsymbol{\Upsilon}_j^T$ and $\psi_{\epsilon j}$ respectively. The prior distributions for the j-th row and diagonal element are different from those in the l-th row and diagonal $j \neq l$. For each of the elements of $\boldsymbol{\Psi_\epsilon}$, we can then specify a Gamma prior

$$\psi_{\epsilon j}^{-1} \sim \text{Gamma}(\alpha_{0\epsilon j}, \beta_{0\epsilon j}).$$

while for the coefficients $\boldsymbol{\Upsilon}_j$, and now given $\psi_{\epsilon j}$, we can specify Normal priors

$$\boldsymbol{\Upsilon}_j | \psi_{\epsilon j} \sim N(\boldsymbol{\Upsilon}_{0j}, \psi_{\epsilon j} \boldsymbol{H}_{0yj})$$

for each of the elements in the matrix of the latent scores $\boldsymbol{\Upsilon}$. In a similar manner, we can specify priors for the covariance matrix $\boldsymbol{\Phi}$ assuming a Wishart distribution.

5.2.4 TTE model with a latent variable as a predictor

The previous sections provided a general framework for fitting SEMs with several latent variables that may be related to other latent variables. However, our interest is in modelling the association of the latent variables with TTE outcomes using observed bio-markers that are assumed to relate to the latent variables. To achieve this, we need to fit SEMs that have a TTE model in the structural part (5.5). Following [56], the effect of the latent variable ω_i on the TTE can be modelled indirectly through the observed variables $\boldsymbol{y_i}$ as follows.

We assume the likelihood for the specified model with the latent covariates is derived from responses that are independent and identically distributed. Further, the response variables, in this case the TTEs are independent, conditional on the latent variables and the observed variables that are assumed to be related to the latent variable.

As before, the full or joint likelihood is just the product of the distribution of the TTE and that of the latent variables. We thus just need to integrate out the latent variables. Let $\boldsymbol{\theta}$ be the vector of model parameters, $\boldsymbol{t}$ be the vector of observed response variables, $\boldsymbol{y}$ be the vector of observed variables that are related to the latent variables, and $\boldsymbol{\omega}$ be the $k \times 1$ vector of latent variables. Then the marginal likelihood can be obtained by integrating out the latent variables as follows

$$L(\boldsymbol{\theta}) = \int f(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\boldsymbol{\mu}_{\boldsymbol{\omega}}, \Sigma_{\boldsymbol{\omega}}) \partial \boldsymbol{\omega} \tag{5.17}$$

Here, $f(.)$ is the conditional pdf for the observed response variables, in this case TTE, $p(.)$ is the multivariate normal density for $\boldsymbol{\omega}$, $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ is the expected value of the latent variables, and $\Sigma_{\boldsymbol{\omega}}$ is the covariance matrix for $\boldsymbol{\omega}$. The vector of unknown parameters $\boldsymbol{\theta}$ includes parameters in both $f(.)$ and $p(.)$.

Assuming a Weibull model for the TTE, we can use the usual model for on the scale parameter

$$\log(\lambda) = \boldsymbol{z} = \exp(\boldsymbol{\beta^T x})$$

with expected value

$$\boldsymbol{\mu}_t = \Gamma(1 + 1/\gamma) \exp(-\boldsymbol{z}/\gamma)$$

to model the effect of the latent variables $\boldsymbol{\omega}$ on $\boldsymbol{t}$. The link function essentially maps the conditional mean of the survival times

$$\boldsymbol{\mu}_t = E(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\omega})$$

to the linear prediction

$$\boldsymbol{z} = \boldsymbol{y}^T \boldsymbol{\beta} + \boldsymbol{y}^T \Upsilon \boldsymbol{\omega} \tag{5.18}$$

where $\boldsymbol{\beta}$ is the vector of the fixed-effect coefficients that relate to the TTE, and Υ is the matrix of factor loadings as previously stated. Ultimately, the likelihood (5.17) can be written as

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\Sigma_{\boldsymbol{\omega}}|}} \int \exp\left\{\log f(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu}_{\boldsymbol{\omega}})^T \Sigma_{\boldsymbol{\omega}}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu}_{\boldsymbol{\omega}})\right\} \partial\boldsymbol{\omega}$$

Further, assuming $\boldsymbol{\omega} \sim N(\mathbf{0}, \boldsymbol{\Phi})$, this becomes

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\Sigma_{\boldsymbol{\omega}}|}} \int \exp\left\{\log f(\boldsymbol{t}|\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\omega}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\omega}\right\} \partial\boldsymbol{\omega}$$

This likelihood is generally not tractable and therefore numerical methods have to be used to estimate the parameters when using ML. In Stata, such models are estimated using Gauss–Hermite quadrature [144]. In MCMC, the independence assumption makes it feasible to specify conditional distributions of the parameters and then use the methods described in Section 2.9.3 to sample from the joint posterior distribution.

5.3 Ensuring model identifiability

Model identifiability is known to be an issue in latent variable models in general [145]. In many SEM applications that are based on ML techniques, the focus is on the sample variance-covariance matrix (S) and how closely it resembles the model implied variance-covariance matrix Σ. In the measurement model (5.1) for instance, the elements of $\Sigma = \boldsymbol{\Upsilon}^T \boldsymbol{\Phi} \boldsymbol{\Upsilon} + \boldsymbol{\Psi}_\epsilon$ are a linear combination of the elements in the matrix of factor loadings $\boldsymbol{\Upsilon}$, the elements in the variance-covariance matrix of the latent variables $\boldsymbol{\Phi}$ and those from the residuals resulting from the measurement equations $\boldsymbol{\Psi}_\epsilon$. All parameters of interest except the intercepts $\boldsymbol{\mu}$ are therefore included in these three matrices that make up Σ. To be able to identify the model, some parameters have to be fixed. Normally, a constant is chosen for either one or more of the elements of the matrix of factor loadings, or for the variances of the latent variables. We thus aim to chose so-called free parameters that are able to vary during the estimation process.

There are two ways of ensuring identification in SEMs, namely:

1. *Unit variance identification*: this is where the variance(s) of the latent variable(s) are fixed at 1 with a mean of 0. This would be more appropriate in the case where interest is on eliciting the posterior distribution of the unknown latent re-growth rate and none of the observed variables can be said to be directly related to the latent growth , and

2. *Unit loading identification*: this is where the factor loading (coefficient) of one of the manifest/observed variables is fixed at 1. This would be appropriate when one of the observed variables which is fixed is known to be directly related to the latent variable.

Both these identification choices will result in the same overall fit although the interpretation of the models will differ between the two choices [146].

5.3.1 Identifiability in a measurement model with 3 observed variables

The number of unknown parameters in the model will determine how many parameters must be fixed to ensure that the model is identifiable. As a general rule, if the latent variable results from k observed variables, the covariance matrix Σ will be of size $k \times k$ with $s = k(k+1)/2$ unknown elements that are to be estimated [147]. These are then compared with the number of elements in the observed covariance matrix S to determine if the model can be identified. If the number of elements in S is less than s, then the model cannot be identified unless further constraints are applied or additional data is available, for example an informative prior distribution in a Bayesian analysis.

As an example, suppose we observe three random variables that associated with a single unobserved latent variable. For this model, assume that $\omega \sim N(0, \phi)$, where the variance of the latent variable, ϕ, is fixed at 1. Then, we can list all the known and unknown elements in the observed and model variance/covariance matrices respectively.

Known from S	**Unknown from Σ**
$Cov(Y_1,\ Y_2)$	$Var(\epsilon_1)$
$Cov(Y_1,\ Y_3)$	$Var(\epsilon_2)$
$Cov(Y_2,\ Y_3)$	$Var(\epsilon_3)$
$Var(Y_1)$	τ_1
$Var(Y_2)$	τ_2
$Var(Y_3)$	τ_3

Table 5.1: Measurable and estimable parameters in a measurement model with 3 observed variables with ϕ fixed at 1. Here, S is the sample covariance matrix, while Σ is the model covariance matrix.

This means that given the observed 3 variables (y_1, y_2, y_3), we have $s = \frac{k(k+1)}{2} = \frac{3(3+1)}{2} = 6$ observed variances/covariances. We also have 6 parameters to estimate given that we have fixed the variance of the latent variable, $\phi = 1$, which allows us to calculate the residual degrees of freedom (df) as $df = s - 6 = 0$. This means our model is identifiable. In general a model is:

1. Over-identified when the residual $df > 0$
2. Just identified or saturated when the residual $df = 0$, and
3. Under-identified when the residual $df < 0$.

The model would also be identifiable if we had fixed one of the factor loadings say, $\tau_1 = 1$ at 1. The models fitted in this analysis will be based on a fixed variance for the latent variable.

5.4 Applications to simulated TTE data

To investigate how well these methods would work with TTE outcomes, a small simulation study was undertaken. The statistical model had the same structure as our intended application. We assumed that 3 observed biomarkers were related to a latent variable that is not observed which was itself a covariate in a TTE model where the simulated TTEs were assumed to come from the two parameter Weibull distribution with scale (λ) and shape (γ) parameters. The latent variable was assumed to have a $N(0,1)$ distribution with the biomarkers related to the latent variable according to the measurement model (5.1), with true intercepts and factor loadings specified for each simulated model. Conditional on the latent variable, the TTE was simulated from the Weibull distribution. The scale parameter in Weibull model was linearly related to the latent variable on the log-scale.

Ways of simulating survival times have been widely discussed [148, 139], so the focus was instead on how well the SEMs with TTE extensions perform in the presence of no censoring, some censoring (around 15%) and considerable censoring ($>$ 30%). The simulated code is shown in Appendix B.

To investigate whether having more observations will lead to better estimates, datasets with $n = 500$ and $n = 5,000$ observations were simulated. Comparisons of how well the models fit the data in each situation were made by looking at the mean (SD) of each parameter over all simulations, the average bias as a percentage, coverage and the confidence interval width for each parameter. The simulations, followed the following steps:

1. Simulate the latent variable, ω_i, as a vector of size $n = 500$ or $n = 5,000$ from the $N(0,1)$ distribution

2. Simulate the measurement errors ϵ_{ij} and specify the μ_j's which are the inter-

cepts, and the τ_j's, which are the factor loadings according to the measurement equations

$$y_{ij} = \mu_j + \tau_j \omega_i + \epsilon_{ij}$$

where $j = 1, 2, 3$ and $i = 1, 2, \ldots, n$, to generate the observed random variables representing the 3 biomarkers.

3. Conditional on the simulated latent variable in step 1, simulate the TTE using the **survsim** package in Stata which is used to simulate survival data from parametric distributions [149]. In survsim, TTEs are simulated by specifying the shape (γ) and scale (λ) parameters for the chosen Weibull TTE model. Covariates can be modelled through the scale parameter and taking the latent variable simulated in step 1 as the covariate on the TTE, we can thus express the log-hazard for the i-th individual as

$$\log(\lambda_i) = \beta_0 + \beta_1 \omega_i$$

where β_0 is the intercept and β_1 is the coefficient for the latent variable as it relates to the hazard of the event. To generate TTEs from the Weibull distribution, we specify ($\gamma \neq 1$) for the shape parameter.

4. To apply censoring, independently simulate TTEs from an Exponential distribution with a fixed scale parameter (λ) and assign the value of the censoring indicator as 1 if this TTE is greater than that in step 3 and 0 otherwise. Different values of λ, β_0 and β_1 were used to generate data sets with varying levels of censoring.

5. Depending on the parameters from step 4 that give the desired level of censoring, fit Weibull TTE models relating the latent variable to the TTE in a SEM, but using only the 3 biomarker variables generated in step 2 using the Stata package **gsem** [150]. The unobserved latent variable was assumed to have a $N(0, 1)$ distribution when fitting these models to ensure model identifiability.

6. Repeat this process for 1,000 replications of the data and work out the mean bias for each estimate as a percentage of the true parameter and the coverage showing in how many of the fitted models, the CI of each estimate included the true parameter.

In this scheme, the models fitted using gsem should produce parameter estimates similar to those specified in the measurement equations of step 2 (μ_j's, τ_j's and the errors ϵ_{ij}) and those from the specified TTE distribution in step 3 (β_0, β_1, and γ). We briefly report on results from simulations under the different censoring levels and sample sizes.

Model performance was assessed following the recommendations by Burton and others [151]. For our parameters of interest $\boldsymbol{\theta}$, we evaluated the resulting estimates $\hat{\boldsymbol{\theta}}$, in terms of bias, coverage and average 95% CI width as shown in Table 5.2.

Evaluation criteria	Formula
Bias as a percentage	$\frac{\bar{\hat{\boldsymbol{\theta}}}-\boldsymbol{\theta}}{\boldsymbol{\theta}} \times 100$
Coverage (COV)	Proportion of times 95% CI for $\hat{\boldsymbol{\theta}}$ includes $\boldsymbol{\theta}$
Average 95% CI length (ACIW)	$\frac{\sum_{j=1}^{J} 3.92 \times \text{SE}(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}})}{J}$

Table 5.2: Performance measures for evaluating model fit. J is the number of simulations, $\bar{\hat{\boldsymbol{\theta}}}$ is the average estimate over all simulations, SE($\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$) is the SE of the estimate of $\boldsymbol{\theta}$ in the j-th dataset, $j = 1, 2, \ldots, J$, while 3.92 represents twice the z-value (1.96) of the 95% CI for the standard normal distribution.

5.4.1 Models with no censoring

We first fit Weibull models assuming each individual has the event to the 1,000 simulated datasets with $n = 500$ and $n = 5,000$ respectively. The estimates and their performance are shown in Table 5.3.

Starting with the smaller dataset, the mean (SE) estimate of the intercept in the TTE model was -0.511 (0.002) which while slightly lower than the set parameter value of -0.5, was reasonably close. Moreover, the bias was $< 1\%$ with a good coverage and a short ACIW. Similarly, the mean of the estimate of the coefficient of the latent variable, β_1 was the same as the target parameter value, with an even lower bias (0.048%). The estimate of the shape was on average slightly higher than the true value but the bias was still less than $< 1\%$. Looking at the SEM part, the estimates of the factor loadings, τ's and intercepts μ's were all close to the true parameter values. The variance estimates in the smaller dataset were slightly smaller than the true values with wider ACIWs for ϵ_1 and ϵ_2 when compared to the other parameters in this model.

When fitted to the bigger datasets, the mean parameter estimates were more precise and in some instances, the same as the true parameters in both the TTE and SEM parts. The SEs for all estimates across all simulations were very small implying the models were able to model the effect of the latent variable on the TTE via the three observed random variables in a SEM that allows for TTE outcomes. The bias as a percentage was even lower in the models fitted to the bigger dataset while the ACIW shrunk to less than 0.300 for even those variance parameters that had wide CIs in the simulations with the smaller dataset. In all models with no censoring, the coverage was around the target 95% for almost all parameters while for some parameters, it tended to exceed nominal 95% target in the larger datasets. However, this was not a big issue and could have been due to chance.

5.4.2 Models with some censoring

To introduce some censoring, we used an independent censoring distribution as detailed in step 4 of the simulation algorithm above. The resulting datasets had on average 17.8%, 95% CI (17.7%, 17.9%) records that did not have the event or were censored. We then proceeded to fit Weibull models as before to the 1,000 datasets of 500 and

No censoring		Parameter		Mean (SE)	Bias %	COV	ACIW
Weibull ($n = 500$)							
	TTE model	β_0	-0.5	-0.514 (0.002)	0.556	97.3	0.251
		β_1	-0.7	-0.700 (0.003)	0.048	94.92	0.315
		γ	1.5	1.509 (0.002)	0.621	100.0	0.266
	SEM part						
		μ_1	1.7	1.699 (0.002)	-0.036	95.4	0.290
		τ_1	0.5	0.497 (0.003)	-0.626	94.2	0.342
		μ_2	3.5	3.502 (0.001)	0.048	94.7	0.165
		τ_2	0.7	0.701 (0.002)	0.131	94.2	0.205
		μ_3	3.0	3.004 (0.003)	0.123	95.5	0.352
		τ_3	1.2	1.194 (0.003)	-0.466	95.6	0.416
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.486 (0.005)	-0.568	94.2	0.646
		ϵ_2	0.4	0.394 (0.002)	-1.575	96.6	0.233
		ϵ_3	2.6	2.593 (0.007)	-0.272	95.7	0.910
Weibull ($n = 5,000$)							
	TTE model	β_0	-0.5	-0.511 (0.001)	-0.027	97.5	0.079
		β_1	-0.7	-0.700 (0.001)	0.050	96.1	0.099
		γ	1.5	1.502 (0.001)	0.100	100.0	0.083
	SEM part						
		μ_1	1.7	1.700 (0.001)	0.021	94.9	0.092
		τ_1	0.5	0.499 (0.001)	-0.179	94.4	0.108
		μ_2	3.5	3.500 (0.000)	-0.012	95.8	0.052
		τ_2	0.7	0.700 (0.000)	0.002	95.8	0.064
		μ_3	3.0	3.001 (0.001)	0.036	94.5	0.111
		τ_3	1.2	1.200 (0.001)	-0.020	94.7	0.131
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.501 (0.002)	0.054	95.6	0.205
		ϵ_2	0.4	0.399 (0.001)	-0.131	95.6	0.072
		ϵ_3	2.6	2.598 (0.002)	-0.080	95.2	0.287

Table 5.3: Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ without censoring.

5,000 respectively to obtain the estimates in Table 5.4.

Again starting with models fitted to the smaller datasets, the parameter estimates in the TTE model were slightly off the target parameter values. However, the bias, as a percentage, of the parameters ranged from -1.5% to 0.7%. The COV for most parameters was around the 95%, while ACIWs were reasonably short for most parameters. A similar trend was observed in the SEM part and also for the variance estimates. The bias was largest for ϵ_2 while the ACIW was biggest for ϵ_3.

Increasing the sample size again resulted in more precise estimates with lower levels of bias, smaller ACIW and mean of the estimates themselves which were accurate to 2 decimal places in general, and SEs that were very small. This illustrates that the SEMs with TTE extensions in the presence of censoring can be used to model TTE outcomes and that larger datasets will result in better and more accurate estimates.

5.4.3 Models with more than 30% censoring

The final check involved fitting the model to simulated data with a lot of censoring. The 1,000 simulated datasets had an average of 39.4% events, 95% CI (39.3%, 39.5%) for both $n = 500$ and $n = 5,000$.

With more than 30% of the observations censored, the estimate of the coefficient of the latent variable, β_1, on the TTE was much lower than the true parameter resulting in a larger negative bias (-34%), Table 5.5. Moreover, COV was at 80% meaning that the CI included the true parameter less than 95% of the time. The same lack of accuracy was observed in the estimates of the factor loadings in the SEM part which had large negative biases of between 28 and 30% with COV around 80%. The variance estimates however, remained reasonably close to the true parameters in general. To check what could have led to these problems, histograms of the distribution of the latent variable as a covariate in the TTE model and in the measurement models were plotted, Figure 5.1. The estimates flipped between negative or positive values which were of the opposite

17.8% censoring		Parameter		Mean (SE)	Bias %	COV	ACIW
Weibull ($n = 500$)							
	TTE model	β_0	-0.5	-0.514 (0.002)	0.712	97.3	0.260
		β_1	-0.7	-0.703 (0.003)	0.383	94.4	0.340
		γ	1.5	1.510 (0.002)	0.681	100.0	0.279
	SEM part						
		μ_1	1.7	1.699 (0.002)	-0.032	95.3	0.290
		τ_1	0.5	0.497 (0.003)	-0.555	94.8	0.344
		μ_2	3.5	3.502 (0.001)	0.057	94.7	0.165
		τ_2	0.7	0.701 (0.002)	0.092	94.3	0.213
		μ_3	3.0	3.005 (0.003)	0.156	95.5	0.352
		τ_3	1.2	1.195 (0.003)	-0.375	95.8	0.428
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.485 (0.005)	-0.604	94.4	0.647
		ϵ_2	0.4	0.394 (0.002)	-1.546	96.3	0.247
		ϵ_3	2.6	2.591 (0.008)	-0.328	95.2	0.940
Weibull ($n = 5,000$)							
	TTE model	β_0	-0.5	-0.511 (0.001)	-0.006	97.6	0.082
		β_1	-0.7	-0.700 (0.001)	-0.025	95.0	0.106
		γ	1.5	1.502 (0.001)	0.104	100.0	0.087
	SEM part						
		μ_1	1.7	1.700 (0.001)	0.021	94.9	0.092
		τ_1	0.5	0.499 (0.001)	-0.186	94.4	0.109
		μ_2	3.5	3.500 (0.000)	-0.012	95.8	0.052
		τ_2	0.7	0.700 (0.001)	-0.006	96.1	0.067
		μ_3	3.0	3.001 (0.001)	0.036	94.5	0.111
		τ_3	1.2	1.200 (0.001)	-0.010	94.6	0.135
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.501 (0.002)	0.055	95.5	0.206
		ϵ_2	0.4	0.400 (0.001)	-0.117	96.1	0.077
		ϵ_3	2.6	2.598 (0.002)	-0.093	94.3	0.295

Table 5.4: Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ with 18% censoring on average.

sign to the true parameter values in some cases, resulting in the bi-modal distributions. In 77/500 (15%) of the datasets, the factor loading of the latent variable on the observed variables, $\tau_j, j = 1, 2, 3$, was negative. This could have been due to chance, but as the bias and COV were averaged over all simulations, it could have led to increase in the bias. Consequently, the bias in the estimate of the effect of the latent variable on the TTE (β_1) would increase as well.

When fitted to the larger datasets, the estimates obtained were reasonably close to the true parameters as expected. Once again all bias, COV, and ACIWs as well as the mean estimates were as expected, almost the same as the true parameters. The problem with the factor loadings and effect of the latent variable on the TTE disappeared with the increase in the sample size.

5.5 Summary

In this chapter, we have introduced SEMs in general and their applications. We have discussed how latent constructs can be modelled using measurement equations that relate observed continuous or binary random variables that are known to be related to the latent variable of interest, for example tumour growth is related to some available biomarkers. We also showed how outcomes of interest can be modelled through structural models to link the latent variable and outcomes of interest. For this thesis, we considered continuous outcomes and then investigated SEMs with TTE outcomes in the structural model. We also discussed how parameters are estimated in SEMs using both ML and Bayesian methods including fixing the variance of the latent variable to ensure identifiability. We then detailed how the full likelihood function and marginal distributions can be determined for each of the parameters by integrating out those parameters that are not of interest.

To illustrate how the SEMs with extensions to include TTE outcomes can be imple-

39% censoring		Parameter		Mean (SE)	Bias %	COV	ACIW
Weibull ($n = 500$)							
	TTE model	β_0	-1.6	-1.622 (0.003)	0.809	97.5	0.382
		β_1	-0.5	-0.329 (0.013)	-34.236	80.2	0.334
		γ	1.5	1.512 (0.002)	0.825	100.0	0.283
	SEM part						
		μ_1	1.7	1.700 (0.002)	-0.022	95.6	0.290
		τ_1	0.5	0.348 (0.012)	-30.404	79.7	0.353
		μ_2	3.5	3.502 (0.001)	0.057	94.7	0.165
		τ_2	0.7	0.498 (0.016)	-28.806	80.6	0.260
		μ_3	3.0	3.004 (0.002)	0.141	95.3	0.352
		τ_3	1.2	0.831 (0.028)	-30.709	79.9	0.492
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.485 (0.005)	-0.596	94.3	0.652
		ϵ_2	0.4	0.387 (0.002)	-3.328	97.0	0.327
		ϵ_3	2.6	2.593 (0.009)	-0.285	94.9	1.107
Weibull ($n = 5,000$)							
	TTE model	β_0	-1.6	-1.611 (0.001)	0.108	98.4	0.120
		β_1	-0.5	-0.501 (0.001)	0.109	95.3	0.105
		γ	1.5	1.502 (0.001)	0.124	100.0	0.088
	SEM part						
		μ_1	1.7	1.700 (0.001)	0.024	94.8	0.092
		τ_1	0.5	0.499 (0.001)	-0.149	95.0	0.112
		μ_2	3.5	3.500 (0.000)	-0.011	95.7	0.052
		τ_2	0.7	0.700 (0.001)	-0.022	95.3	0.080
		μ_3	3.0	3.001 (0.001)	0.033	94.5	0.111
		τ_3	1.2	1.200 (0.001)	0.035	94.8	0.154
	Variances						
		$Var(\eta)$	1.0				
		ϵ_1	2.5	2.501 (0.002)	0.048	95.7	0.207
		ϵ_2	0.4	0.400 (0.001)	-0.096	96.2	0.099
		ϵ_3	2.6	2.596 (0.003)	-0.164	93.6	0.346

Table 5.5: Mean and empirical SEs for each parameter in Weibull TTE models with a single latent variable as a covariate. The SEM part gives estimates of the factor loadings τ's, intercepts μ's and variances ϵ's. The bias as a percentage, coverage (COV) and average CI width (ACIW) are also reported. Results are presented for $n = 500$ and $N = 5,000$ with 39% censoring on average.

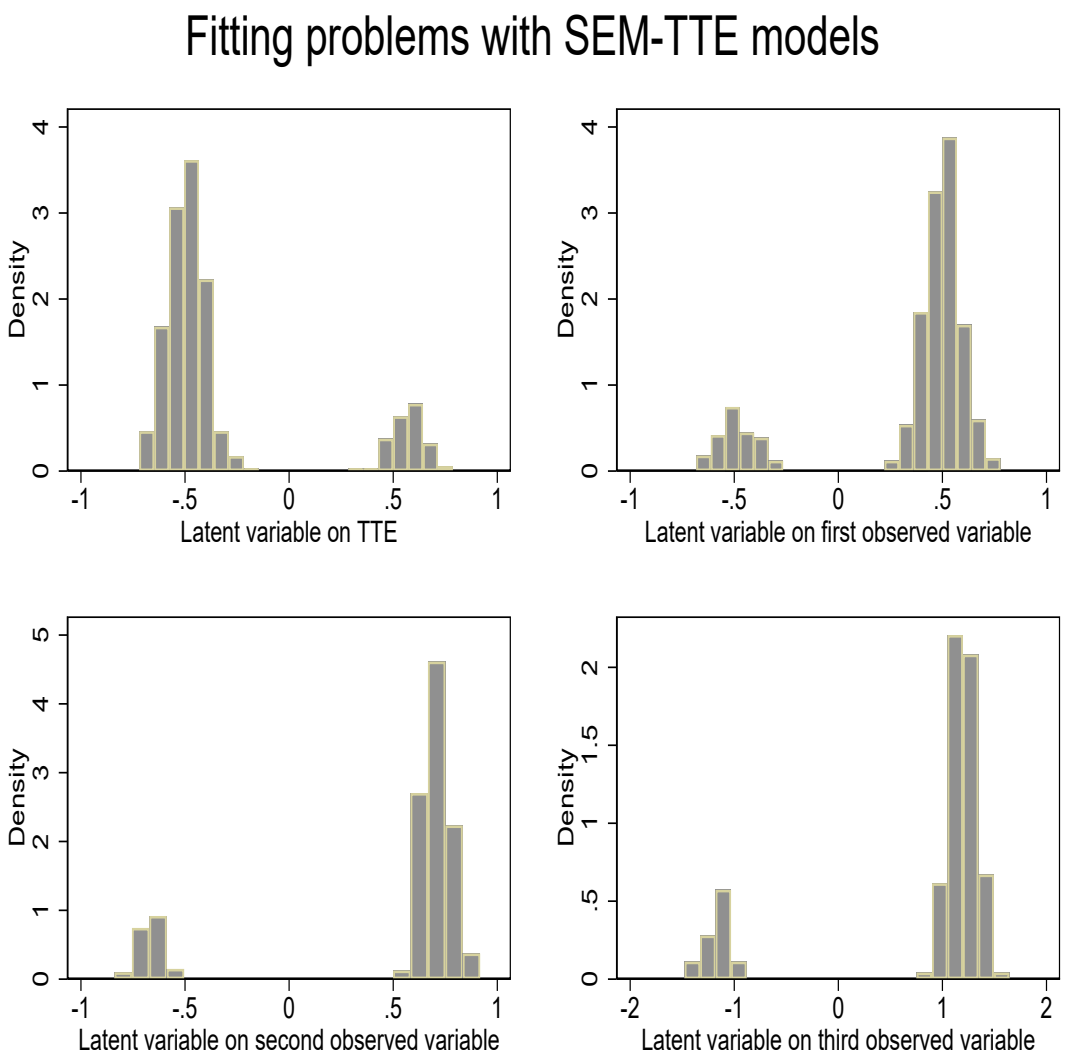


Figure 5.1: Histograms depicting the distribution of the latent variable as a covariate on the TTE model and in the measurement models from simulated datasets with $n = 500$. The bi-modal distributions illustrate problems with accuracy in some of the simulations.

mented, we fitted various models with TTE outcomes in the structural model to simulated TTE data from Weibull models with various levels of censoring, by fitting models to small and large datasets. The simulations showed that it was generally possible to obtain precise estimates in models with no or some censoring $<$ 20%. The precision of

the estimates improved with increasing sample size in most cases. Detailed guidance has been provided on the ideal size of datasets to which SEMs can be fitted, with a general rule of thumb stating that an average of 10 cases per unknown parameter are required [152]. Our models had on average, 9 parameters to be estimates which with data from a minimum of 500 rows made it possible to obtain reasonable estimates.

The simulation showed that with increased censoring, it was hard to obtain accurate estimates of the effect of the latent variable on both the TTE and as it relates to the 3 observed variables via the measurement equations. In smaller datasets, accuracy was influenced by a few extreme results. Increasing the sample size resulted in better estimates in general. However, it is worth bearing in mind that with ML techniques, there is need to verify that the global maximum has been reached. In SEMs fitted via ML, a simple way of checking model validity is to check that the latent variable is significantly associated with both the outcome in the structural (TTE model) and the observed variables (markers) as a way of verifying that the data supports the proposed model.

The discussion so far has focused on SEMs with TTE outcomes that are normally used when analysing survival data. However, we are also interested in how we could fit cure rate models that have latent variables as covariates. We thus apply further extensions to these SEMs to model the effect of tumour growth on the TTR and OS in Myeloma as well on the proportion that will be cured following treatment. However, we also found in chapter 5 that these latent variable models can be subject to identifiability problems. Therefore, if we were to include another latent variable (the marker of whether an individual is cured or not), it is likely to cause further computational problems. To explore this issue, a preliminary simulation study was undertaken (results not included) which showed cure rate models with latent covariates can be fitted within a Bayesian framework with only mildly informative priors and we illustrate their application in the next chapter.

Chapter 6

Using SEMs to model the role of unobserved tumour growth on TTE outcomes: application to Myeloma data

6.1 Introduction

In this chapter, we illustrate how SEMs can be used to model the role of tumour growth rate following treatment on TTR and OS using the methods developed in Chapter 5. We also use these models to estimate the proportion cured following treatment using cure rate models with applications to the Myeloma dataset, further extending the current methodology that is limited to standard TTE outcomes.

The previous applications to the two example datasets investigated the role of fixed covariates such as log-RD, age, sex, paraprotein, beta2 and albumin on TTR and OS using the well-known parametric and semi-parametric TTE models. Cure rate models were then used to assess how these covariates are associated with the proportion who will never relapse and the proportion whose OS returns to that of age-sex matched individuals in the general population after treatment.

In these analyses, important predictors of these outcomes were available from the

data sets except the tumour growth rate after treatment, which is another important predictor of TTR, OS and possibly the probability of a patient being cured. As the growth rate was not directly measured in the Myeloma dataset, we can use the available variables that are known to measure the propensity for growth in a SEM framework. In the Myeloma dataset, paraprotein, albumin and beta2 are associated with TTR and OS. They can also each be used as surrogates of how fast the tumour is likely to grow.

The cytogenetic aberrations could be used within a SEM with binary observed variables to model the effect of growth on the TTE outcomes. However, most of the patients did not have these cytogenetic aberrations available in the Myeloma dataset, Table 3.2. For the rest of this thesis, we will assume that paraprotein, beta2 and albumin are the observed variables, that together are manifestations or markers for the latent variable tumour growth rate which is assumed to be a normally distributed continuous random variable with a mean of 0 and variance 1.

Using this realised measure of the tumour growth rate, the measurement model will relate these observed markers to the latent variable which will itself be a covariate in TTR and OS models, and cure rate models. When modelling the TTR and OS, we fit Cox PH and Weibull models with the latent variable growth as a fixed covariate. For the cure rate models, we again consider mixture and PTC models with a Weibull distribution for the TTR and OS respectively for those that are not cured. We finally include log-RD in all models to investigate how both these important prognostic factors relate to the outcomes. As the RD percentage was measured after treatment while paraprotein, beta2 and albumin were measured at baseline, we assume, for purposes of this analysis, that values of these three variables were unchanged from those at baseline at the time of RD percentage measurement. We next briefly describe the dataset used in these analyses.

6.1.1 Brief description of the data for modelling the effect of tumour growth on the TTE outcomes in Myeloma

As some patients did not have records of some of the markers for growth, we only considered those patients with paraprotein, beta2 and albumin. The final dataset had data from 243 patients. With respect to TTR, there were 187 (77%) patients who relapsed during the follow-up period with median TTR 2.1 years (95% CI: 1.78 - 2.38 years) and 94 (38.7%) deaths during follow-up, Figure 6.1. The K-M estimates of survivor functions with TTR and OS show that the patterns are similar to the full data set. We thus proceed to perform the CC analysis on this sub-set of the data.

As SEMs assume multivariate normality, we took logarithms of paraprotein, beta2 and albumin in order to make these variables approximately normally distributed. Since some patients had values of zero for some biomarkers after treatment, we take a logarithm of 1 plus each of paraprotein, beta2 and albumin before fitting the models. The distributions of each of the observed variables and scatter plots for pairs of biomarkers are shown in Figure 6.2. The log-transformed beta2 and albumin were approximately normally distributed, while for paraprotein, those without paraprotein in their blood (paraprotein = 0) dominated. The correlations between these 3 measures were small in general especially between paraprotein and beta2. For a well specified SEM, the observed variables should be considerably correlated. As this analysis is mainly exploratory, we proceed to model the effect on tumour growth on the TTE outcomes in the Myeloma dataset.

From the resulting log-transformed biomarkers, there were no obvious differences in the median log-paraprotein and log-beta2 between those who relapsed and those who were censored at the end of the follow-up, while patients who relapsed had lower log-albumin values on average. For OS, those who died had higher levels of log-paraprotein and lower log-albumin compared to those who did not, while there were no obvious

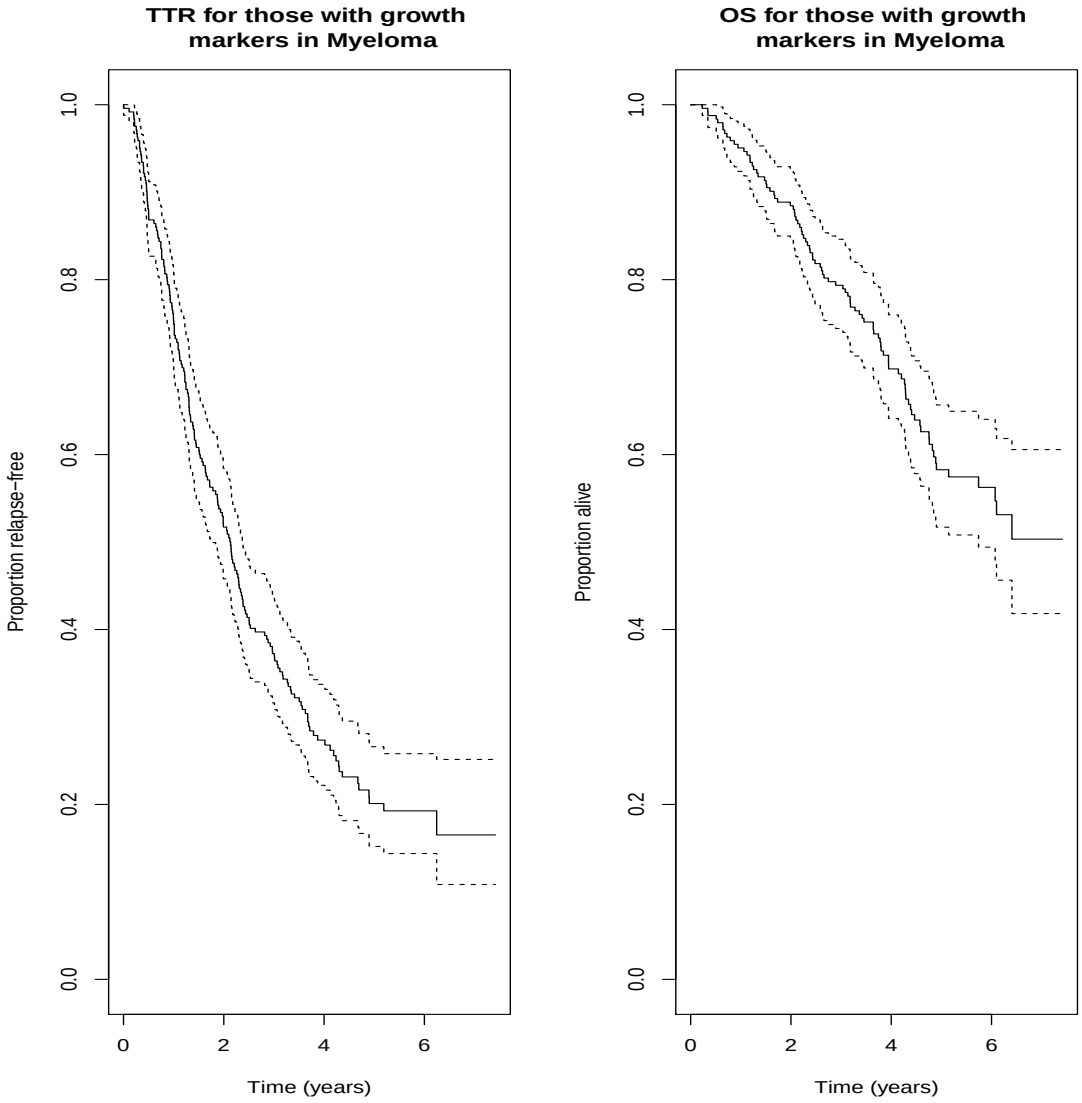


Figure 6.1: K-M plots for the TTR and OS for 243 Myeloma patients with paraprotein, beta2 and albumin data. Dotted lines represent 95% CI limits.

differences with regards to log-beta2, Figure 6.3. Once again, due to differences in follow-up time, we do not focus on these differences in our narrative.

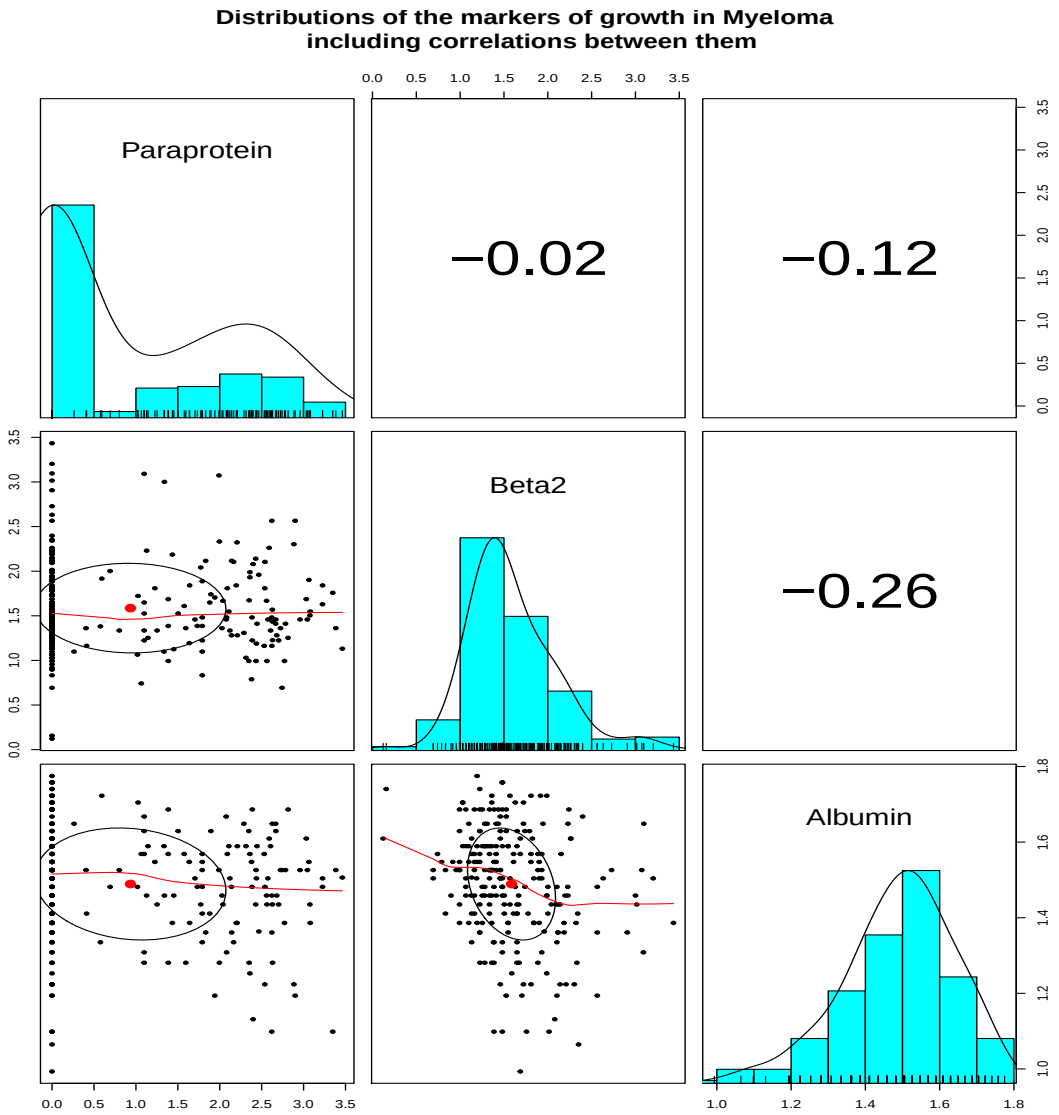


Figure 6.2: Pairs plot showing the distributions of the log-transformed biomarkers as well as the correlation between them

6.2 TTE models with growth and log-RD in Myeloma

We considered models for TTR/OS, first with growth rate as a latent covariate on its own, related through the three observed bio-markers, and growth rate together with

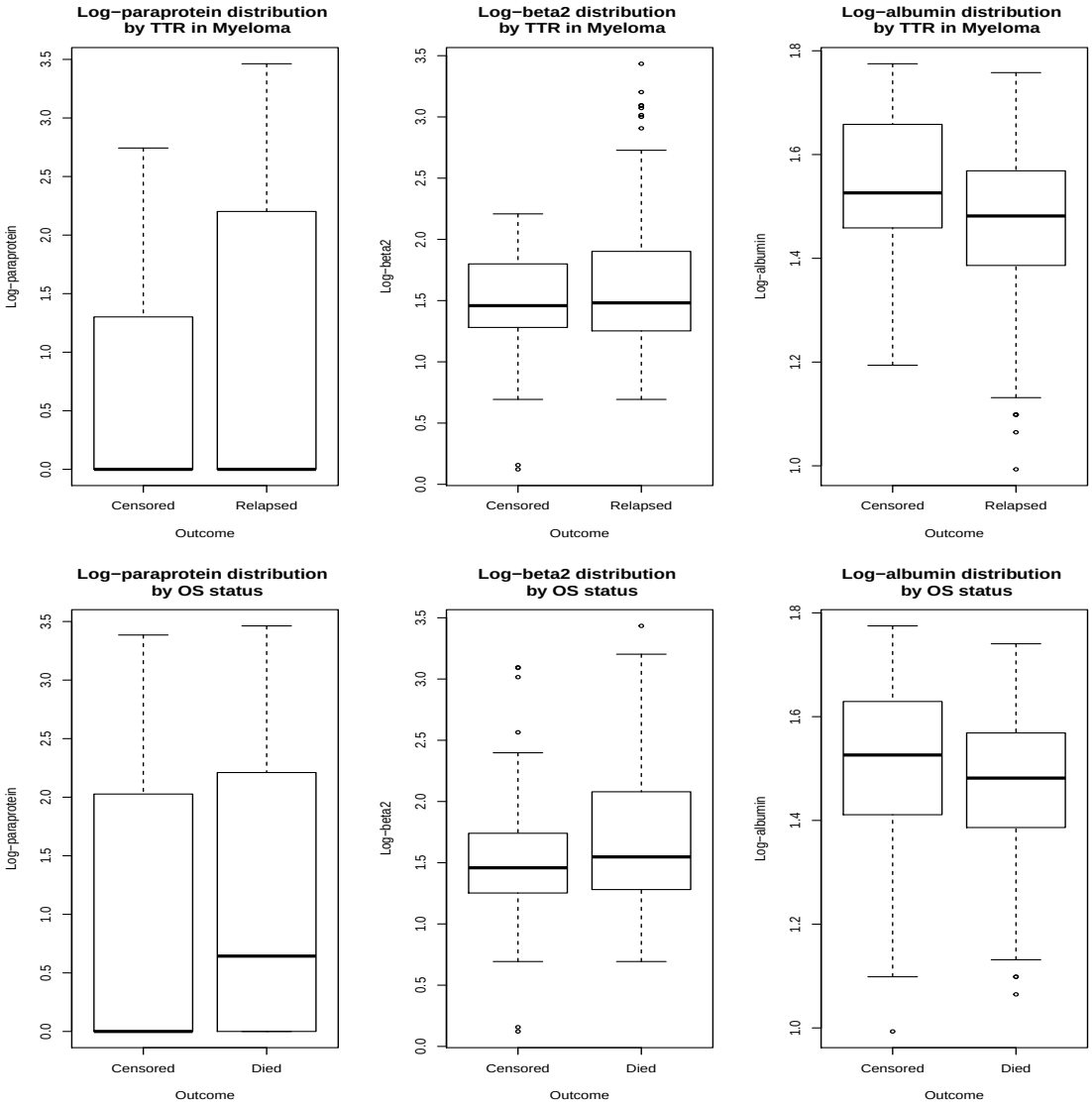


Figure 6.3: Boxplots showing log-transformed paraprotein, beta2 and albumin by relapse and survival status at the end of the study in Myeloma

log-RD in multivariable models for TTR/OS as shown in Figure 6.4. In these models, we assumed that the growth rate at recruitment was the same as that at the time of RD measurement. The growth rate and log-RD can thus both be treated as fixed covariates in the TTR, OS and cure rate models.

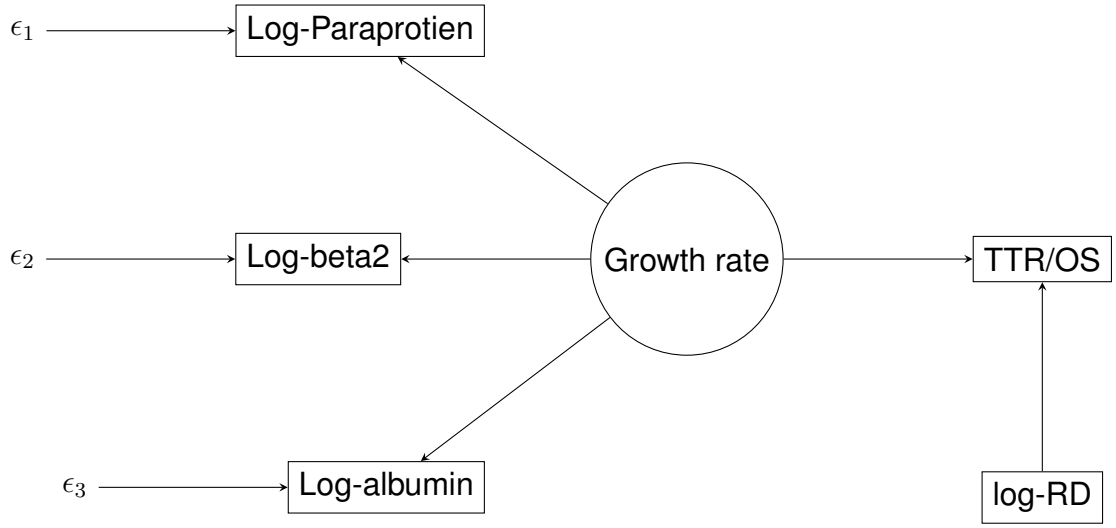


Figure 6.4: SEM model for TTR or OS with tumour growth rate and RD

6.2.1 Measurement model for the growth rate

The growth rate was realised from a single factor measurement model with the three observed variables on the log-scale; paraprotein, beta2 and albumin as markers. The measurement models for the i-th individual, $i = 1, \ldots, n$ can be expressed as

$$\boldsymbol{y_i} = \boldsymbol{\mu} + \boldsymbol{\tau}\omega_i + \boldsymbol{\epsilon_i} \tag{6.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$ are 3×1 vectors of the intercepts and factor loadings in the linear models relating the observed random variables log-paraprotein, log-beta2 and log-albumin respectively, to the latent growth rate and $\boldsymbol{\epsilon_i}$ is a patient-specific 3×3 diagonal matrix of the errors relating to the linear models of the biomarkers and latent tumour growth and $\boldsymbol{y_i}$ is a vector of 3 biomarker measurements for individual $i = 1, ..., n$. The full measurement model assuming a variance of 1 for the growth rate had the following parameters to be estimated:

Intercepts

- μ_1: the intercept in the linear model of the effect of tumour growth on log-paraprotein

- μ_2: the intercept in the linear model of the effect of tumour growth on log-beta2
- μ_3: the intercept in the linear model of the effect of tumour growth on log-albumin

Factor loadings

- τ_1: the coefficient of growth in the linear model relating tumour growth to log-paraprotein
- τ_2: the coefficient of growth in the linear model relating tumour growth to log-beta2
- τ_3: the coefficient of growth in the linear model relating tumour growth to log-albumin

Variances

- ϵ_1: the variance of log-paraprotein
- ϵ_2: the variance of log-beta2
- ϵ_3: the variance of log-albumin

6.2.2 Prior distributions and constraints

When using a Bayesian approach, it is also necessary to specify prior distributions for model parameters in BUGS language. Assuming independence, we can specify priors for the latent variable and the TTE model by considering only the form of the conditional distributions of each of the unknowns in turn. For the measurement model

$$y_{ij} = \mu_j + \tau_j \omega_i + \epsilon_{ij},$$

we specify vague Normal priors for the intercept $\mu_j \sim N(0, 100)$ and factor loading $\tau_j \sim N(0, 100), j = 1, 2, 3$ and ω_i represents the tumour growth rate for the i-th individual. For the variance components, we specify Inverse-Gamma priors for the diagonal elements

for Ψ_ϵ, $\psi_j^{-1} \sim InvGamma(\rho_0, \rho_1)$ with $\rho_0 = \rho_1 = 1$ for $j = 1, 2, 3$. We constrain the variance of the ω_i's to be 1 with a mean of 0.

For the TTR and OS, we specify a Weibull model with the growth rate, ω_i, modelled on the scale using $\lambda_i = \exp(\beta_0 + \beta_1\omega_i)$. We then specify Normal priors for β_0 and β_1 as $N(0, 100)$ and for the shape γ we specify a $Gamma(1, 1)$ prior. We also consider a Cox PH model

$$h(t_i) = h_0(t_i)\lambda_i$$

as before, where now $\lambda_i = \exp(\beta_1\omega_i)$ and $h_0(t_i)$ is the unspecified baseline hazard as before. Priors for the baseline hazard are specified as in Section 2.9.3, while the usual $N(0, 100)$ prior is used for the hazard ratio, β_1.

To fit cure rate models, we use the logistic model as before, but, with the latent growth rate, ω_i, as the covariate on π

$$\text{log}\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1\omega_i,$$

and again specify Normal priors for α_0 and α_1, $N(0, 100)$.

Finally, to model the association of both the growth and log-RD with the TTR and OS in cure rate models, we specify the multivariable model for λ as

$$\lambda_i = \exp(\beta_0 + \beta_1\omega_i + \beta_2 x_i)$$

and the model for π as

$$\text{log}\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1\omega_i + \alpha_2 z_i$$

where ω_i is the growth rate and x_i represents log-RD as a covariate on the TTR/OS and z_i is the log-RD as a covariate in the model for π. For these extra parameters, we again

specify $N(0, 100)$ priors.

Using this set-up, we fit Cox PH and Weibull models for TTR and OS with a SEM component to generate draws from the posterior distribution of the tumour growth rate from the measurement model based on log-paraprotein, log-beta2 and log-albumin using MCMC. Where possible we also fit models for TTR and OS assuming a Weibull distribution for the observed times in Stata using generalised structural equation model (**gsem**), a package that enables one to fit SEMs with generalised outcomes including a parametric survival model (in this case), and one or more latent variables as covariates in the TTE outcome [150]. We also fit cure rate models with growth as a covariate on both π and TTR/OS as well as multivariable models with both growth and log-RD as covariates using MCMC. We next present results from each of these different models.

To assess goodness of fit, we used graphical means such as trace and density plots to determine whether the MCMC chain of the conditional posterior distribution for each individual parameter had converged. In cure rate models, graphical checks of convergence were made using Gelman plots [153], which show how the PSFR changes through the iterations. These plots compare the within and between chain variability and convergence is said to have been achieved if the PSFR converges to 1 after some iterations. The rule of thumb is that a PSFR of 1.1 indicates that the chains have converged to the target posterior distribution.

6.3 The effect of tumour growth on the TTR in Myeloma

Results from Bayesian Cox PH, Bayesian Weibull and a ML based Weibull model with the latent variable tumour re-growth rate as a covariate on TTR in Myeloma are shown in Table 6.1. The Bayesian Weibull and Cox PH models were fitted in OpenBUGS using the code Appendix C.1 and C.2 respectively, while the Weibull ML model was fitted in Stata using gsem. In all 3 models, the growth rate was significantly associated with TTR

for these patients log-HR (95% CrI or CI) 1.463 (0.671, 2.497) in the Cox PH model, 1.385 (0.066, 2.705) in the Weibull ML model, and 1.454 (0.858, 2.347) in the Bayesian Weibull model. The estimated hazard ratios from these three models suggest that a SD increase in growth rate was associated with an increase in the hazard of relapsing of 4.32, 3.99 and 4.28, respectively. All 3 models showed that a fast growing tumour was predictive of a shorter TTR. The Bayesian Weibull model had the narrowest CrI among the 3 models in the growth parameter on TTR. For the measurement part, the growth rate was positively associated with both log-paraprotein and log-beta2 and negatively associated with log-albumin. Estimates of the intercepts, factor loading and variance components were similar in all 3 models. The association of the growth rate with both the observed variables and the TTR illustrates that the proposed model underlying the latent construct, growth rate, was supported by the observed data. These results illustrate that the effect of the unobserved growth rate on the TTR could be indirectly modelled through the three observed variables in a SEM set-up.

As a check for goodness of fit, trace and density plots of the posterior distribution of log-HR for growth on the TTR in the Bayesian Cox PH and Bayesian Weibull models were used, Figure 6.5. Poor mixing was evident especially in the Bayesian Cox PH model while the Weibull seemed to achieve convergence to a unique mean of the posterior density. This could be remedied by running longer chains in both models. The apparent struggles with fitting in the Cox PH model can be related to the number of events. In this analysis, 187 individuals relapsed and since the counting process used to fit the Bayesian Cox PH model compares outcomes at each event time, the Cox model in effect had less data from which to estimate the parameters in the TTR model. The Weibull ML model fitted in Stata is well established and would thus naturally be the model of choice. The Bayesian models nonetheless, perform as well as the ML equivalent in estimating the effect of growth on the TTR.

	Bayesian Cox PH	**Weibull ML**	**Bayesian Weibull**
Parameter	Estimate (95 % CrI)	Estimate (95 % CI)	Estimate (95 % CrI)
TTR model (λ)			
Intercept (β_0)	-	-1.748 (-2.488, -1.008)	-1.785 (-2.378, -1.365)
Growth (β_1)	1.463 (0.671, 2.497)	1.385 (0.066, 2.705)	1.454 (0.858, 2.347)
Shape (γ)	-	1.677 (0.985, 2.855)	1.733 (1.316, 2.368)
SEM part			
Paraprotein			
Intercept (μ_1)	0.939 (0.797, 1.081)	0.936 (0.793, 1.079)	0.943 (0.803, 1.088)
Growth (τ_1)	0.382 (0.182, 0.611)	0.366 (0.178, 0.554)	0.375 (0.197, 0.576)
Beta2			
Intercept (μ_2)	1.586 (1.524, 1.649)	1.586 (1.523, 1.649)	1.587 (1.525, 1.651)
Growth (τ_2)	0.093 (0.005, 0.184)	0.121 (0.023, 0.219)	0.106 (0.024, 0.194)
Albumin			
Intercept (μ_3)	1.489 (1.459, 1.518)	1.489 (1.471, 1.508)	1.488 (1.464, 1.511)
Growth (τ_3)	-0.035 (-0.068, -0.003)	-0.053 (-0.084, -0.022)	-0.035 (-0.062, -0.009)
Variances			
$Var(\omega)$	1 (fixed)	1 (fixed)	1 (fixed)
Var(Paraprotein) (ϵ_1)	1.150 (0.908, 1.408)	1.160 (0.954, 1.410)	1.166 (0.953, 1.420)
Var(Beta2) (ϵ_2)	0.270 (0.226, 0.325)	0.237 (0.195, 0.288)	0.253 (0.209, 0.307)
Var(Albumin) (ϵ_3)	0.053 (0.044, 0.063)	0.019 (0.015, 0.024)	0.033 (0.028, 0.039)
LL/DIC	3,056	-813.6	7,815

Table 6.1: Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for TTR fitted to $n = 243$ Myeloma patients with tumour growth rate as a covariate. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

6.4 The effect of tumour growth and log-RD on the TTR in Myeloma

When adjusted for log-RD, a faster growth rate was again associated with a shorter TTR, Table 6.2. In these adjusted models, effect of the growth rate slightly decreased in

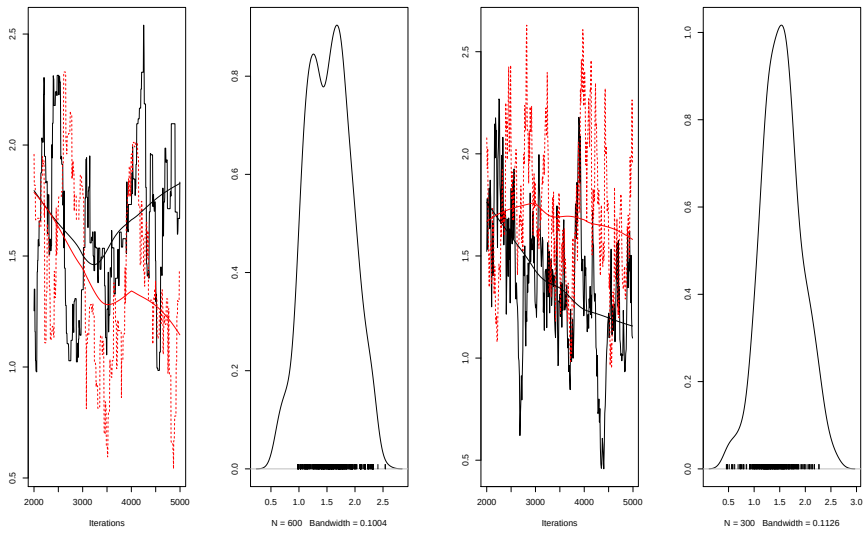


Figure 6.5: Trace and density plots of the effect growth on the TTR from the Bayesian Cox PH model (A) and from the Bayesian Weibull model (B) in the Myeloma dataset.

all three models; log-HR 1.463 vs 1.347 in the Cox PH model, log-HR: 1.385 vs 1.198 in the Weibull ML model and log-HR: 1.454 vs 1.385 in the Bayesian Weibull model when the log-RD was taken into account. This would be expected because log-RD is also associated with TTR and therefore some of the risk of a relapse was attributable to log-RD. Each increase on the log-scale of the RD percentage was associated with a 0.538, 0.280 and 0.303 increase in the hazard of relapsing on the log-scale in the Cox PH, Weibull ML and Bayesian models respectively. The effect in both Weibull models was not significant. By comparison, the log-HR in the Cox PH model for TTR adjusted for paraprotein, beta2 and albumin fitted to the same number of patients as part of the sensitivity analysis (243) was 0.207 (-0.088, 0.502) Section 3.2.10 which was also not significant.

Looking at the measurement part, the growth rate was again positively associated with both log-paraprotein and log-beta2 and negatively associated with log-albumin as

expected. However, for log-beta2 in the SEM part of the Bayesian Cox PH model, the CrI for the estimate of the effect of growth rate included zero. This means the proposed SEM model was not well supported by the observed variables in the Cox PH mainly due to convergence problems similar to those seen in the univariable Cox PH with 5,000 iterations, and also because we used non-informative priors in setting up the MCMC. The power to detect the effect of both log-RD and growth on TTR was low because of the small sample size and number of events. Again, running longer chains would remedy this problem. Other than that, all other estimates of the intercepts, factor loadings and variance components were similar in the ML and Bayesian Weibull models, meaning the hypothesised latent construct representing growth rate was supported by the observed data. By having a CrI that included a 0, there was no evidence supporting the association of the growth and paraprotein and therefore the existence of the latent variable growth unlike in the two other models.

Finally looking at the DIC and LL in the models in Table 6.1 and Table 6.2, adding log-RD does not affect the fit of the models in both the Weibull ML (change in LL: -812.7 - (-813.6) = 0.9) and Bayesian Weibull (change in DIC: 7812 - 7803= 9) models. For the Cox PH model, the change in DIC was bigger (3045 - 3030 = 15), implying the simpler model was adequate for these data. However, we retained both covariates in all models as interest was in modelling the effect of both of them on the TTR while using them in model adjustment. Again, we would recommend using the well established Weibull ML model fitted in Stata to model the effect of growth and log-RD on the TTR while pointing out that the Bayesian Weibull model performs equally well.

6.5 The effect of tumour growth on OS in Myeloma

To model the effect of tumour growth on OS, the Bayesian Cox PH, Weibull ML and Bayesian Weibull models were used to estimate log-HRs and parameters from the mea-

	Bayesian Cox PH	**Weibull ML**	**Bayesian Weibull**
Parameter	Estimate (95 % CrI)	Estimate (95 % CI)	Estimate (95 % CrI)
TTR model (λ)			
Intercept (β_0)	-	-1.477 (-2.110, -0.843)	-1.540 (-2.074, -1.073)
Growth (β_1)	1.347 (0.209, 3.188)	1.198 (0.234, 2.161)	1.385 (0.760, 2.143)
Log-RD (β_2)	0.538 (0.143, 1.245)	0.280 (-0.133, 0.694)	0.303 (-0.125, 0.750)
Shape (γ)	-	1.561 (1.037, 2.349)	1.683 (1.250, 2.239)
SEM part			
Paraprotein			
Intercept (μ_1)	0.947 (0.765, 1.086)	0.935 (0.792, 1.078)	0.928 (0.781, 1.081)
Growth (τ_1)	0.386 (0.097, 0.933)	0.339 (0.141, 0.537)	0.332 (0.109, 0.552)
Beta2			
Intercept (μ_2)	1.592 (1.528, 1.657)	1.586 (1.523, 1.649)	1.586 (1.524, 1.649)
Growth (τ_2)	0.095 (-0.019, 0.197)	0.136 (0.031, 0.240)	0.109 (0.023, 0.206)
Albumin			
Intercept (μ_3)	1.487 (1.461, 1.517)	1.489 (1.471, 1.508)	1.490 (1.468, 1.512)
Growth (τ_3)	-0.035 (-0.069, -0.002)	-0.057 (-0.091, -0.024)	-0.042 (-0.070, -0.017)
Variances			
$Var(\omega)$	1 (fixed)	1 (fixed)	1 (fixed)
Var(Paraprotein) (ϵ_1)	1.087 (0.416, 1.371)	1.178 (0.970, 1.432)	1.189 (0.957, 1.450)
Var(Beta2) (ϵ_2)	0.260 (0.217, 0.315)	0.233 (0.190, 0.286)	0.248 (0.205, 0.296)
Var(Albumin) (ϵ_3)	0.051 (0.042, 0.060)	0.019 (0.015, 0.024)	0.028 (0.024, 0.034)
LL/DIC	3,045	-812.7	7,812

Table 6.2: Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for TTR fitted to $n = 243$ Myeloma patients with tumour growth rate adjusted for log-RD as covariates. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) for all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

surement models, Table 6.3.

A faster tumour growth rate was predictive of a shorter time to death in all models. For instance the estimated log-HR was 0.589 in the Cox PH model, 0.527 in the Weibull ML model and even higher (0.844) in the Bayesian Weibull model. The CrI for the

log-HR estimated in the Bayesian model was very wide, suggesting a lack of precision which could have resulted from convergence problems.

In the measurement models, the estimates for the intercepts and variances in the Weibull ML and Bayesian Weibull were similar to those in the TTR models. However, the association of growth and log-paraprotein in all models as well as growth and beta2 in the Bayesian Cox PH and Bayesian Weibull models was not significant as the CrIs and CIs included a 0. Further, in the Bayesian Cox PH model, the association of growth and albumin was positive, contrary to the negative association observed in the other models. This suggests that the relationship between these biomarkers and OS, via the latent variable, was weak. Moreover, in the SEM part for the Cox PH model, the estimate of the variance for log-albumin, was different to those from the other two models. The problems with these models might have stemmed from attempting to fit complex models with limited to a small dataset with fewer events in the follow-up period. In this case, the Weibull ML model provided better estimates than the other models despite the obvious problems with the available data which did not support the existence of the latent construct representing growth through the 3 observed covariates.

6.6 The effect of tumour growth and log-RD on OS in Myeloma

Adjusting for the log-RD did not lead to a change in the effect of the growth rate on the TTR in the Weibull ML model, while it resulted in a decrease in the log-HR from 0.844 to 0.673 in the Bayesian Weibull model. The estimate of the effect of growth on the TTR in the Cox PH model was very big (5.702) and different from the other two models, Table 6.6. Ignoring the estimated log-HRs from the Cox PH model which were not plausible, the estimated log-HR for log-RD was 0.292 in both the Weibull ML and Bayesian Weibull models when the growth rate was taken into account. This effect was bigger than the estimate (0.207) obtained when log-RD was adjusted for paraprotein,

	Bayesian Cox PH	**Weibull ML**	**Bayesian Weibull**
Parameter	Estimate (95 % CrI)	Estimate (95 % CI)	Estimate (95 % CrI)
OS model (λ)			
Intercept (β_0)	-	-3.138 (-3.653, -2.623)	-3.364 (-4.414, -2.713)
Growth (β_1)	0.589 (0.168, 1.246)	0.527 (0.112, 0.942)	0.844 (0.202, 1.811)
Shape (γ)	-	1.456 (1.203, 1.761)	1.544 (1.205, 2.0432)
SEM part			
Paraprotein			
Intercept (μ_1)	0.936 (0.796, 1.075)	0.935 (0.792, 1.078)	0.937 (0.796, 1.084)
Growth (τ_1)	0.215 (-0.024, 0.460)	0.182 (-0.051, 0.416)	0.267 (-0.022, 0.800)
Beta2			
Intercept (μ_2)	1.584 (1.525, 1.648)	1.586 (1.523, 1.649)	1.586 (1.523, 1.647)
Growth (τ_2)	0.199 (-0.090, 0.309)	0.232 (0.109, 0.355)	0.160 (-0.012, 0.307)
Albumin			
Intercept (μ_3)	3.479 (3.399, 3.557)	1.489 (1.471, 1.508)	1.489 (1.468, 1.510)
Growth (τ_3)	0.330 (0.175, 0.472)	-0.079 (-0.119, -0.038)	-0.040 (-0.070, -0.009)
Variances			
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.248 (1.030, 1.518)	1.260 (1.046, 1.517)	1.210 (0.629, 1.496)
Var (Beta2) (ϵ_2)	0.219 (0.167, 0.276)	0.198 (0.145, 0.269)	0.229 (0.165, 0.289)
Var(Albumin) (ϵ_3)	0.305 (0.196, 0.407)	0.016 (0.010, 0.024)	0.029 (0.024, 0.034)
LL/DIC	2,662	-731.1	7,757

Table 6.3: Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for OS fitted to $n = 243$ Myeloma patients with tumour growth rate as a covariate. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

beta2 and albumin in the standard Cox PH model fitted to data from the same number of patients in Table 3.8, Section 3.2.10. In both cases, the effect of log-RD was not significant.

The measurement model again had challenges in establishing the relationship be-

tween the observed log-paraprotein and the latent growth rate (CrI for τ_1 included a zero in all three models). In addition to this, the association of growth and albumin could not be established in the Cox PH model. Other than that, the other estimates of the factor loadings and intercepts as well as variances were similar to those in the other models for the Weibull ML and Bayesian Weibull models. The addition of log-RD as a predictor of OS led to an increase in the DIC from 7,757 to 7,762 in the Bayesian Weibull model, an increase in the LL from -731.1 to -730.2 in the Weibull ML model, and an unexpected decrease in the DIC in the Bayesian Cox PH model from 2,662 to 1,819 which was reflective of the implausibly large estimate of the log-HR. The addition of log-RD in the model for OS did not result in a much higher DIC, and since interest was on the effect of growth and log-RD on the OS, we would use the multivariable model rather than just including growth on its own. Again, we would use the well established Weibull ML as the model of choice. Clearly, the association of paraprotein and growth could not be established in all 3 models, suggesting that the data did not support the existence of the latent construct representing growth, through the three observed variables paraprotein, beta2 and albumin. Trace plots of the coefficient of growth (beta1) and log-RD (beta2) in the Bayesian Cox PH model for OS were used to assess the fit of the Bayesian Cox PH model, Figure 6.6. The trace plots of the coefficients of growth and log-RD in the model for OS clearly showed that the two chains did not mix well to arrive at the same target distribution over the 5,000 iterations. The chains explored different parts of the sample space across iterations in general which could account for the log-HR estimates which were very different from the estimates from both Weibull models. While all models had problems in the SEM part, the problems in the Bayesian Cox PH model were more pronounced, making it clear that the existence of the latent growth rate through the three observed variables was not supported by the observed data in this model.

	Bayesian Cox PH	**Weibull ML**	**Bayesian Weibull**
Parameter	Estimate (95 % CrI)	Estimate (95 % CI)	Estimate (95 % CrI)
OS model (λ)			
Intercept (β_0)	-	-2.952 (-3.518, -2.386)	-3.004 (-3.718, -2.424)
Growth (β_1)	5.702 (4.845, 6.554)	0.527 (0.127, 0.926)	0.673 (0.200, 1.378)
Log-RD (β_2)	6.707 (5.611, 7.888)	0.292 (-0.134, 0.718)	0.292 (-0.122, 0.742)
Shape (γ)	-	1.453 (1.201, 1.757)	1.460 (1.187, 1.791)
SEM part			
Paraprotein			
Intercept (μ_1)	0.887 (0.740, 1.039)	0.935 (0.792, 1.078)	0.938 (0.797, 1.084)
Growth (τ_1)	-0.145 (-0.325, 0.045)	0.164 (-0.072, 0.400)	0.143 (-0.117, 0.414)
Beta2			
Intercept (μ_2)	1.614 (1.542, 1.684)	1.586 (1.523, 1.649)	1.583 (1.516, 1.649)
Growth (τ_2)	0.080 (0.001, 0.164)	0.239 (0.109, 0.368)	0.219 (0.073, 0.364)
Albumin			
Intercept (μ_3)	1.484 (1.453, 1.516)	1.489 (1.471, 1.508)	1.490 (1.469, 1.511)
Growth (τ_3)	-0.018 (-0.051, 0.017)	-0.078 (-0.119, -0.037)	-0.043 (-0.075, -0.008)
Variances			
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.238 (1.041, 1.481)	1.266 (1.053, 1.523)	1.272 (1.041, 1.535)
Var(Beta2) (ϵ_2)	0.265 (0.223, 0.315)	0.195 (0.139, 0.272)	0.210 (0.134, 0.275)
Var(Albumin) (ϵ_3)	0.052 (0.044, 0.061)	0.016 (0.011, 0.024)	0.028 (0.024, 0.034)
LL/DIC	1,819	-730.2	7,762

Table 6.4: Estimates from Cox PH, Weibull ML, and Bayesian Weibull models for OS fitted to $n = 243$ Myeloma patients with tumour growth rate adjusted for log-RD as covariates. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) for all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

6.7 Modelling the effect of tumour growth in cure rate models for TTR

The role of tumour growth in the probability of not relapsing following treatment was investigated using cure rate models extended to include the latent variable tumour growth

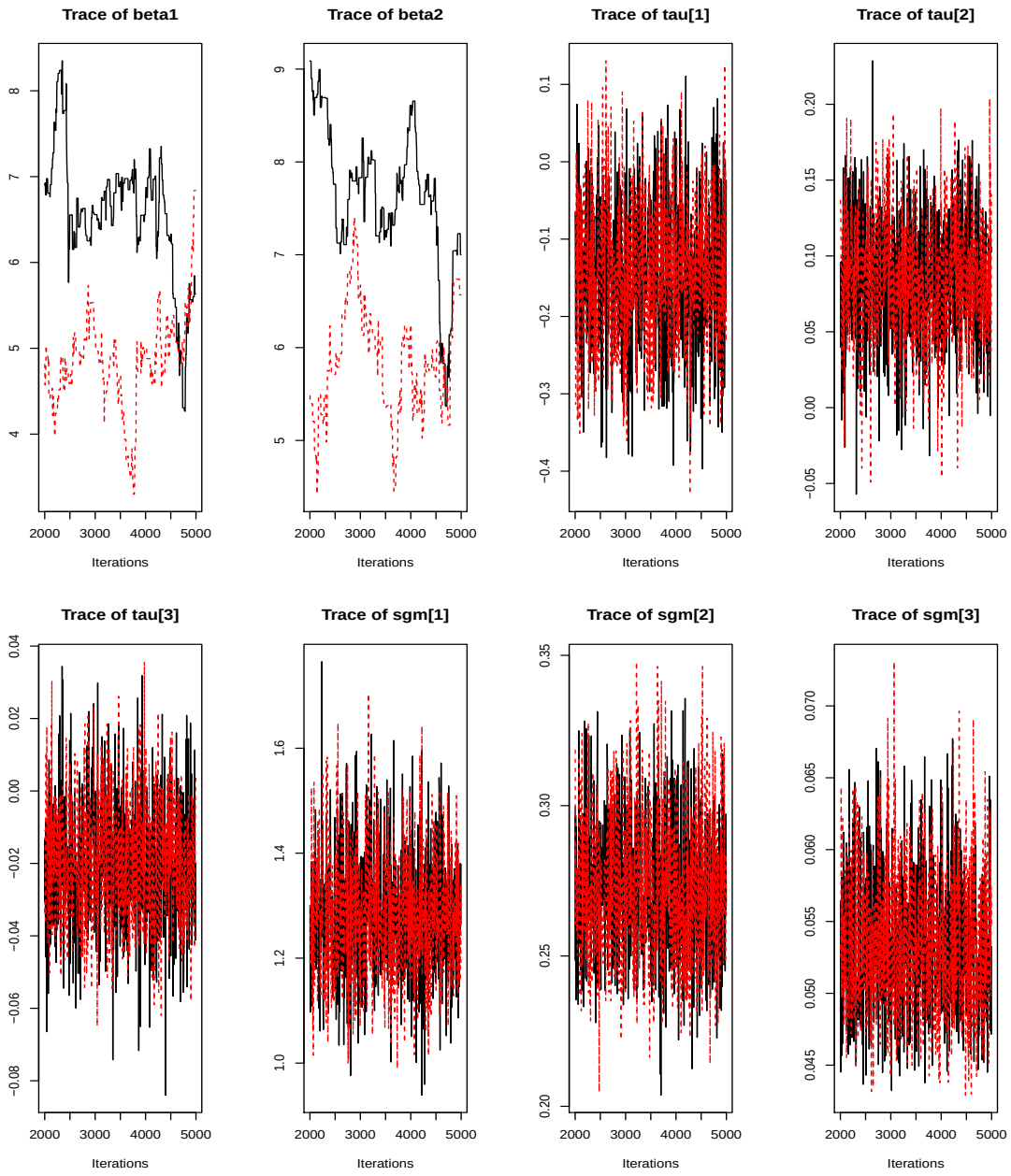


Figure 6.6: Trace plots for the log-ORs for growth (beta1) and log-RD (beta2) as well as factor loadings tau[1] - tau[3] and variances for paraprotein (var[1]), beta2 (var[2]) and albumin (var[3]) from the Bayesian Cox PH model for OS with growth and log-RD as covariates in Myeloma.

rate as a fixed covariate on both π and on the scale in the TTR model, Table 6.5. The OpenBUGS code for fitting these models is shown in Appendix D.1.

In the logistic model for π, each SD increase in the growth rate was associated with a lower probability of being cured in both the mixture (log-OR estimate -1.813) and PTC models (log-OR estimate -1.372). The CrIs in both models for the cured proportion included a 0 and were very wide. For the TTR on the other hand, those who had faster growing tumours were at a higher risk of relapsing in both models. Each SD increase in the growth rate was associated with a 1.289 (0.766, 1.896) increase in the mixture and a 0.911 (0.081, 1.853) increase in the PTC model in the log-hazard of relapsing. In both models, the growth rate was clearly associated with the TTR but not with the probability of being cured following treatment. Comparatively, the effect of growth on the TTR was smaller in the mixture and PTC models when compared to that in the standard TTR model in Table 6.1.

Histograms showing the distribution of the posterior means of the growth rate for each individual in the dataset, as well as the probability of being cured for each individual against the growth rate from each model are shown in Figure 6.7. In both models, there was a clear inverse relationship between the estimated tumour growth rate and the probability of being cured following treatment. Those within -1.5 SDs of the mean growth had the highest probability of being cured, while those with aggressive tumours, for example 1 SD above the mean growth rate, had the lowest probability of being cured.

From the measurement parts in both models, the distribution of the tumour growth was centred around 0, as expected. However, it is worth noting that these distributions of the growth rate did not span the whole range, [-3, 3], associated with the $N(0, 1)$ distribution that was assumed for the growth rate. This could have been due to problems with model fit, or the small sample size or a combination of both. Estimates from the SEM part were similar to those from the Cox PH and Weibull models for TTR and OS, implying the models were fitted using the same underlying latent construct representing

growth rate. Moreover, all the factor loadings were significant, with sensible variance estimates in both models. The growth rate was positively associated with paraprotein and beta2, and had a negative relationship with albumin as expected.

These models show that faster growing tumours were predictive of lower chances of being cured and a high risk of relapsing among those not cured. In terms of fit, the mixture model seemed to fit the data better than the PTC model (DIC 7,810 versus 7,848). Further looking at the Gelman plots from the mixture model, the PSRFs for the factor loadings in the measurement model, tau[2] and tau[3], the coefficient for growth in the model for π and on the scale parameter in the TTR model, alpha1, and beta1 respectively all tended to move towards 1 after around 4,000 iterations, Figure 6.9. However, the Gelman plot for the factor loading associated with paraprotein, tau[1] and its intercept mu[1] depicted PSRFs above 1.1 across the iterations. Other than that, the conditional posterior distributions of the parameters in the mixture model converged quicker based on the between and within chain variance being almost the same. On the other hand, the Gelman plots of the factor loadings, parameters on π and λ as well as the intercepts in the PTC model were less stable as the PSRF tended to oscillate over the iterations meaning the two chains used in the MCMC were slower in converging to the posterior, Figure 6.8. However, the PSRF still fell within the target value of 1.1.

6.8 Modelling the effect of tumour growth and log-RD in cure rate models for TTR

Estimates of log-ORs in the model for π and log-HRs in the TTR model following treatment for growth rate and log-RD from Bayesian Weibull mixture and PTC models are shown in Table 6.6.

When adjusted for log-RD, the estimated log-ORs for growth in the model for the cured proportion were -1.550 and -1.482 in the mixture and PTC models respectively,

	Weibull mixture	**Weibull PTC**
Parameter	Estimate (95 % CrI)	Estimate (95 % CrI)
Cure model (π)		
Intercept (α_0)	-3.738 (-6.334, -1.916)	-3.973 (-7.338, -1.962)
Growth (α_1)	-1.813 (-4.862, 0.578)	-1.372 (-4.408, 0.957)
TTR model (λ)		
Intercept (β_0)	-1.664 (-2.041, -1.270)	-2.884 (-3.687, -2.101)
Growth (β_1)	1.289 (0.766, 1.896)	0.911 (0.081, 1.853)
Shape (γ)	1.727 (1.363, 2.145)	1.734 (1.412, 2.218)
SEM part		
Paraprotein		
Intercept (μ_1)	0.924 (0.778, 1.070)	0.931 (0.799, 1.034)
Growth (τ_1)	0.368 (0.194, 0.546)	0.402 (0.198, 0.565)
Beta2		
Intercept (μ_2)	1.580 (1.527, 1.633)	1.586 (1.526, 1.652)
Growth (τ_2)	0.117 (0.030, 0.198)	0.114 (0.024, 0.204)
Albumin		
Intercept (μ_3)	1.490 (1.465, 1.510)	1.490 (1.470, 1.511)
Growth (τ_3)	-0.045 (-0.074, -0.016)	-0.044 (-0.072, -0.016)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.164 (0.964, 1.404)	1.147 (0.932, 1.403)
Var(Beta2) (ϵ_2)	0.247 (0.206, 0.295)	0.246 (0.201, 0.298)
Var(Albumin) (ϵ_3)	0.028 (0.023, 0.035)	0.028 (0.023, 0.034)
DIC	7,810	7,848

Table 6.5: Bayesian Weibull mixture and PTC models for the proportion that will never relapse and the TTR with growth rate as covariate on both π and λ fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-OR as a coefficient for growth (α_1) are presented. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

implying each SD increase in the growth was associated with a decrease in the log-odds of being cured in both models by 1.550 and 1.482 respectively. However, the CrIs

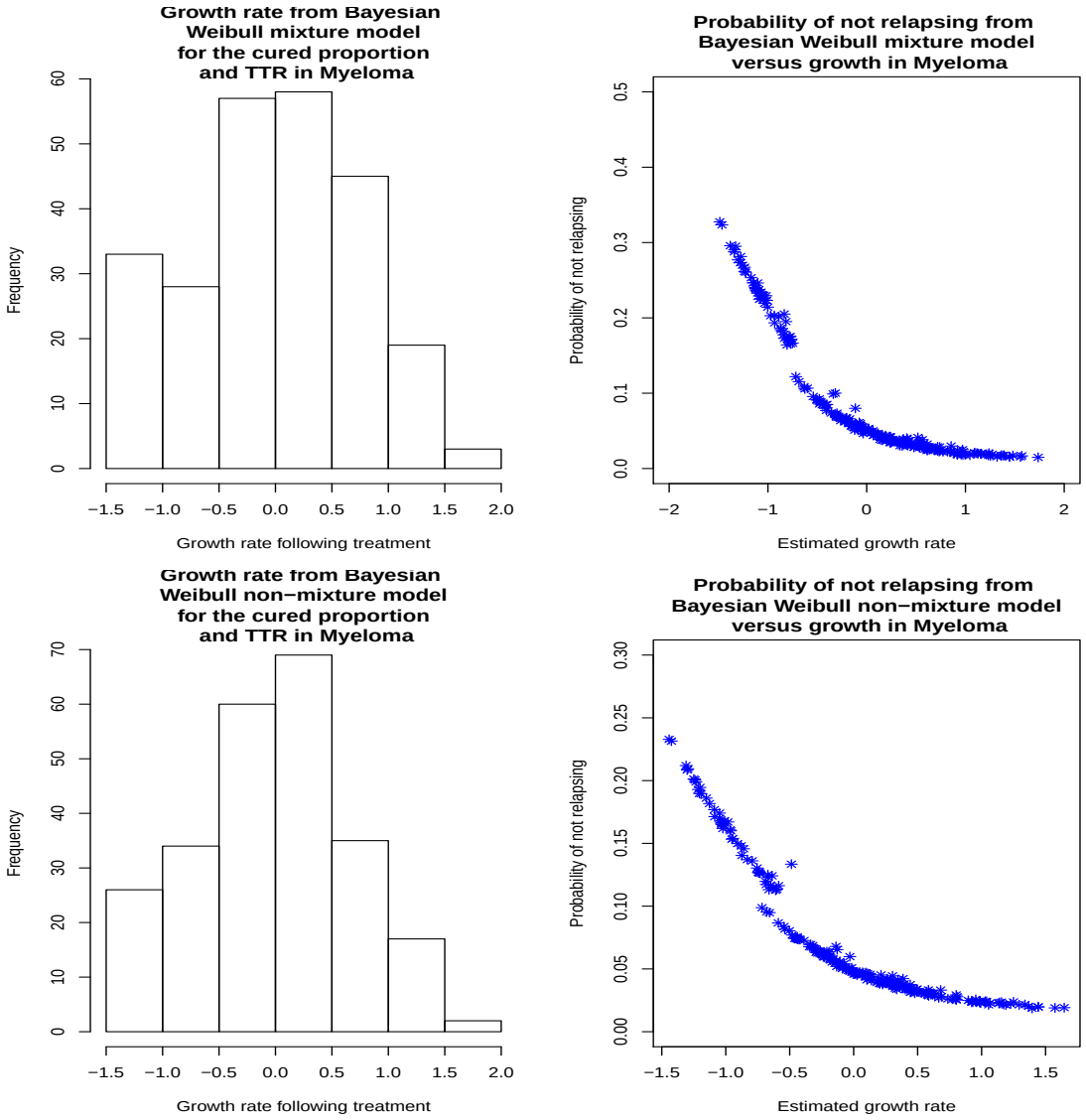


Figure 6.7: Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of being cured versus the growth rate from Weibull mixture model and PTC model in the Myeloma dataset

of the estimates in both models included a 0. Comparatively, the log-OR for growth as the only covariate in the mixture model for π was bigger at -1.813, while in the PTC model it was smaller, -1.372. Once again a higher disease burden was associated with a lower probability of being cured following treatment, log-OR -0.625 and -0.353 in the

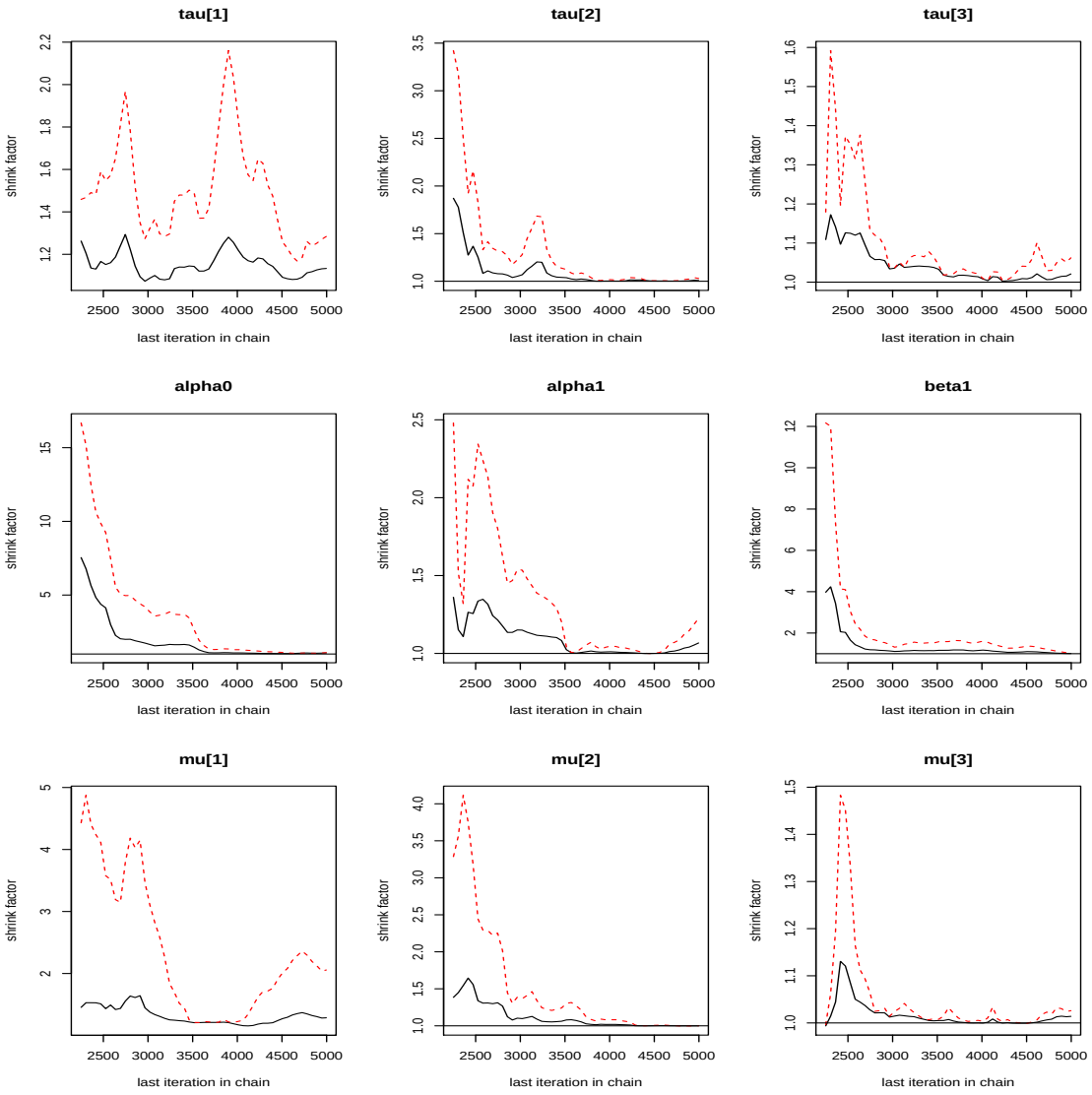


Figure 6.8: Gelman plots showing the PSRF values across iterations in the mixture model for the cured proportion in the Myeloma dataset. Plotted on the first row are the factor loadings, coefficients of growth on π (alpha1) and λ (beta1) are on the second row, while the third row shows Gelman plots for the intercepts in the measurement models.

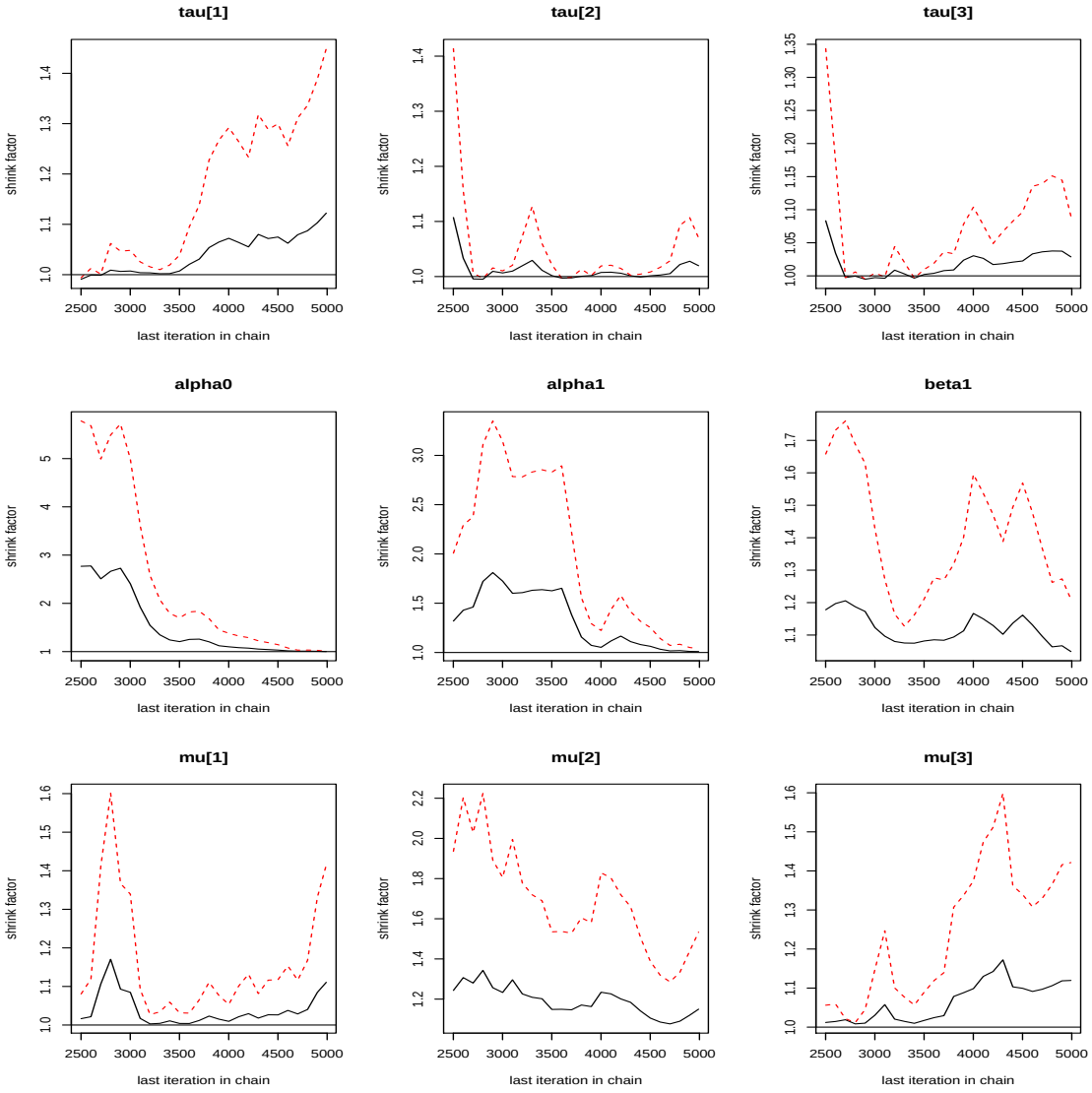


Figure 6.9: Gelman plots showing the PSRF values across iterations in the PTC model for the cured proportion in the Myeloma dataset. Plotted on the first row are the factor loadings, coefficients of growth on π and λ are on the second row, while the third row shows Gelman plots for the intercepts in the measurement models.

mixture and PTC models respectively. The estimated log-OR in the mixture model was bigger than that in the PTC model, although the CrIs for both estimates included a zero. In contrast, log-RD on its own, was a predictor of the probability of not relapsing again after treatment in the cure rate models of Chapter 4. It is worth bearing in mind that the models in Chapter 4 were fitted to a bigger dataset.

For the TTR, each SD increase in the growth rate was associated with an increase in the log-hazard of relapsing; log-HR 1.280 and 0.916 in the mixture and PTC models respectively. The effect in the mixture model was different from 0 based on the CrI while the PTC model estimate CrI included a zero. Adjusting for the growth rate, an increase in the RD percentage on the log-scale accounted for some of the increase in the hazard of relapsing among those who were not cured following treatment; log-HR 0.173 and 0.031 in the mixture and PTC models respectively. This was in contrast to the negative, though not significant, effect of the log-RD on the TTR in Table 4.2. On the other hand, the effect of the log-RD adjusted for the growth rate was similar in direction, but less in magnitude when compared to the effect of the log-RD on TTR adjusted for paraprotein, beta2 and albumin in the Cox PH model for TTR in Table 3.8 when fitted to the same number of patients. This was to be expected as the models in Chapter 3 did not consider the possibility of a cure. The risk of a relapse increased rapidly with time in both the mixture and the PTC models with estimates of the shape parameter greater than those from the cure rate models of Chapter 4. Again, this could be explained by the growth rate driving the TTR for those not cured having taken into account the proportion who might have been cured by the treatment.

For the measurement parts of the model relating the growth rate and its related observed variables, estimates of the intercepts, factor loadings and variance estimates in both models were similar to those from the previous models, suggesting that the data supported the existence of the latent growth rate observed through the three variables. All CrIs for the estimates of the intercepts and coefficients of growth in the measure-

ment equations did not include a zero. Moreover, each SD increase in the growth rate was associated with an increase in the log-transformed paraprotein and beta2 and a decrease in albumin as expected. The distribution of the growth rate adjusted for log-RD in the mixture and the PTC models for TTR was approximately normally distributed, Figure 6.10. Similar to when growth rate was fitted as the only covariate on both the cured proportion and TTR, the probability of being cured decreased with increasing growth rate. In both the mixture and PTC models, the relationship was more scattered in the adjusted models than when growth was the only covariate as shown in Figure 6.7.

Between the two models, the mixture model seemed more effective in modelling the role of growth and log-RD on both the probability of being cured and the TTR as the DIC was smaller (7,818) when compared to the PTC model which had a DIC of 7,838. An examination of the PSRFs across iterations in the mixture model showed that the chains tended towards the posterior for the coefficient of log-RD (alpha2) in the model for π and also in the TTR model (beta2). The Gelman plot for the coefficient of growth in the TTR model (beta1) also showed quick convergence towards the posterior. For the coefficient of growth in the model for π (alpha1), the PSRF tended to oscillate more meaning there were still some differences between the two chains even towards 5,000 iterations. This could explain the wide CrI around the estimated log-OR for the growth rate in the mixture model. There were blips in the plots for the first and second factor loading for log-beta2, tau[2], Figure 6.11. In the PTC model on the other hand, the PSRFs were more varied for most of the parameters in both the cure, TTR and SEM models indicating that this model would need more iterations to reach convergence than the mixture model, Figure 6.12. In this model, only the Gelman plot for the coefficient of log-RD in the model for π (alpha2), and in the TTR model, beta2, showed good convergence. Based on this comparison, we would again recommend using the mixture model to estimate the proportion who are cured in Myeloma as it performed better than

the PTC model in general despite both models struggling with convergence after 5,000 iterations.

	Weibull mixture	**Weibull PTC**
Parameter	Estimate (95 % CrI)	Estimate (95 % CrI)
Cure model (π)		
Intercept (α_0)	-3.992 (-7.226, -2.222)	-4.198 (-7.086, -2.433)
Growth (α_1)	-1.550 (-4.371, 0.419)	-1.482 (-3.312, 1.029)
Log-RD (α_2)	-0.625 (-2.088, 1.446)	-0.353 (-2.005, 2.260)
TTR model (λ)		
Intercept (β_0)	-1.504 (-1.953, -1.010)	-2.885 (-3.763, -2.112)
Growth (β_1)	1.280 (0.691, 1.935)	0.916 (-0.063, 1.888)
Log-RD (β_2)	0.173 (-0.352, 0.644)	0.031 (-0.670, 0.780)
Shape (γ)	1.743 (1.368, 2.159)	1.828 (1.473, 2.267)
SEM part		
Paraprotein		
Intercept (μ_1)	0.949 (0.823, 1.096)	0.929 (0.803, 1.052)
Growth (τ_1)	0.346 (0.136, 0.498)	0.363 (0.162, 0.577)
Beta2		
Intercept (μ_2)	1.591 (1.523, 1.654)	1.584 (1.516, 1.654)
Growth (τ_2)	0.119 (0.037, 0.219)	0.113 (0.028, 0.211)
Albumin		
Intercept (μ_3)	1.489 (1.469, 1.510)	1.489 (1.466, 1.510)
Growth (τ_3)	-0.045 (-0.071, -0.021)	-0.043 (-0.071, -0.016)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.180 (0.966, 1.428)	1.176 (0.945, 1.444)
Var(Beta2) (ϵ_2)	0.247 (0.204, 0.297)	0.247 (0.205, 0.297)
Var(Albumin) (ϵ_3)	0.028 (0.023, 0.034)	0.028 (0.023, 0.034)
DIC	7,818	7,838

Table 6.6: Bayesian Weibull mixture and PTC models for the proportion that will never relapse and the TTR with growth rate and log-RD as covariates on both π and λ fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-ORs as coefficients for growth (α_1) and log-RD (α_2) respectively, are presented. In the TTR model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) respectively from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

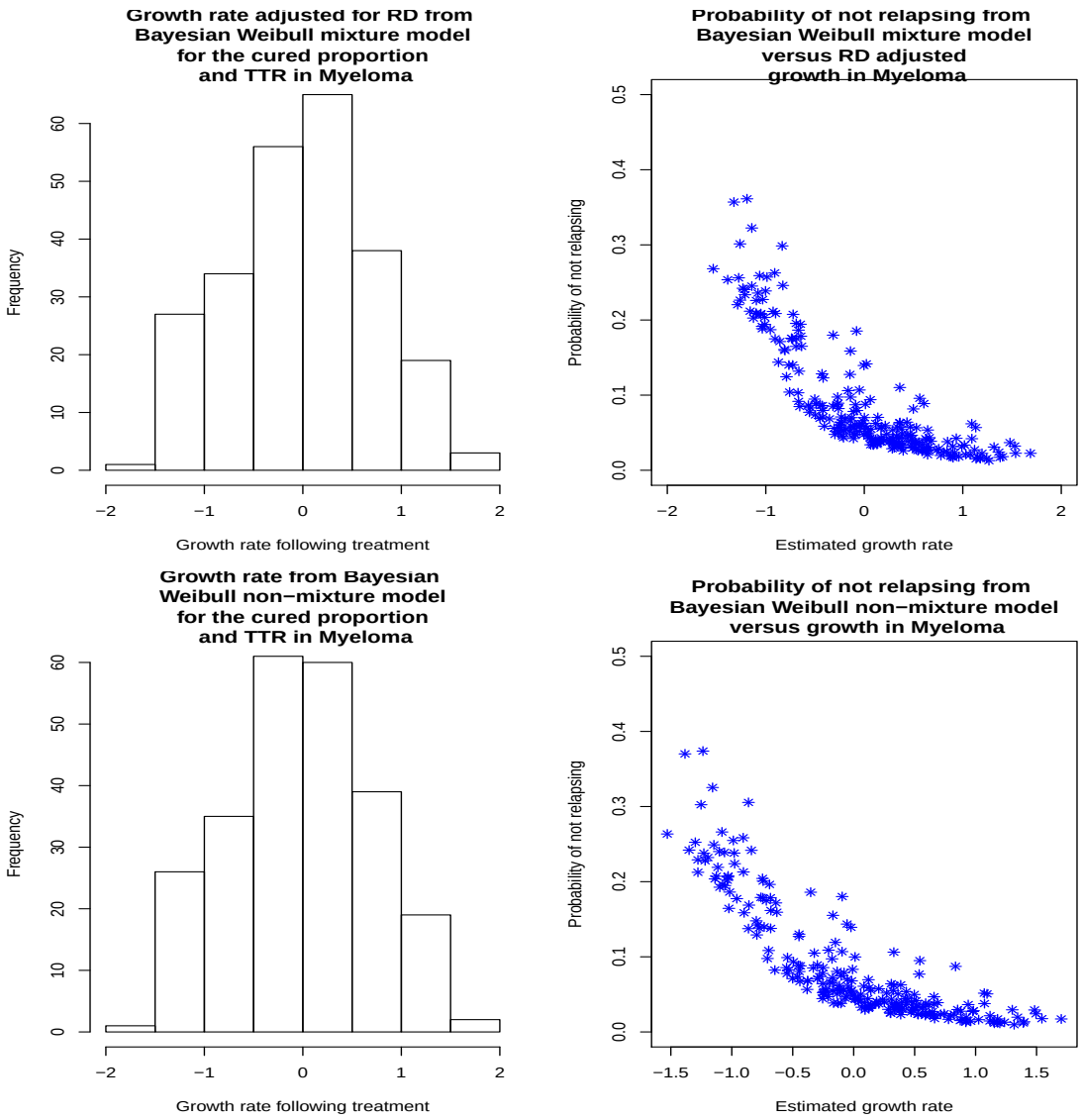


Figure 6.10: Histograms depicting the posterior distribution of the growth rate adjusted for log-RD as well as scatter plots of the probability of being cured versus the log-RD adjusted growth rate from Weibull mixture and PTC models in the Myeloma dataset

6.9 Tumour growth rate as a covariate in population mixture and PTC models

The role of growth on the proportion whose OS returns to that of the general population was investigated using Bayesian mixture and PTC models, Table 6.7. The models were

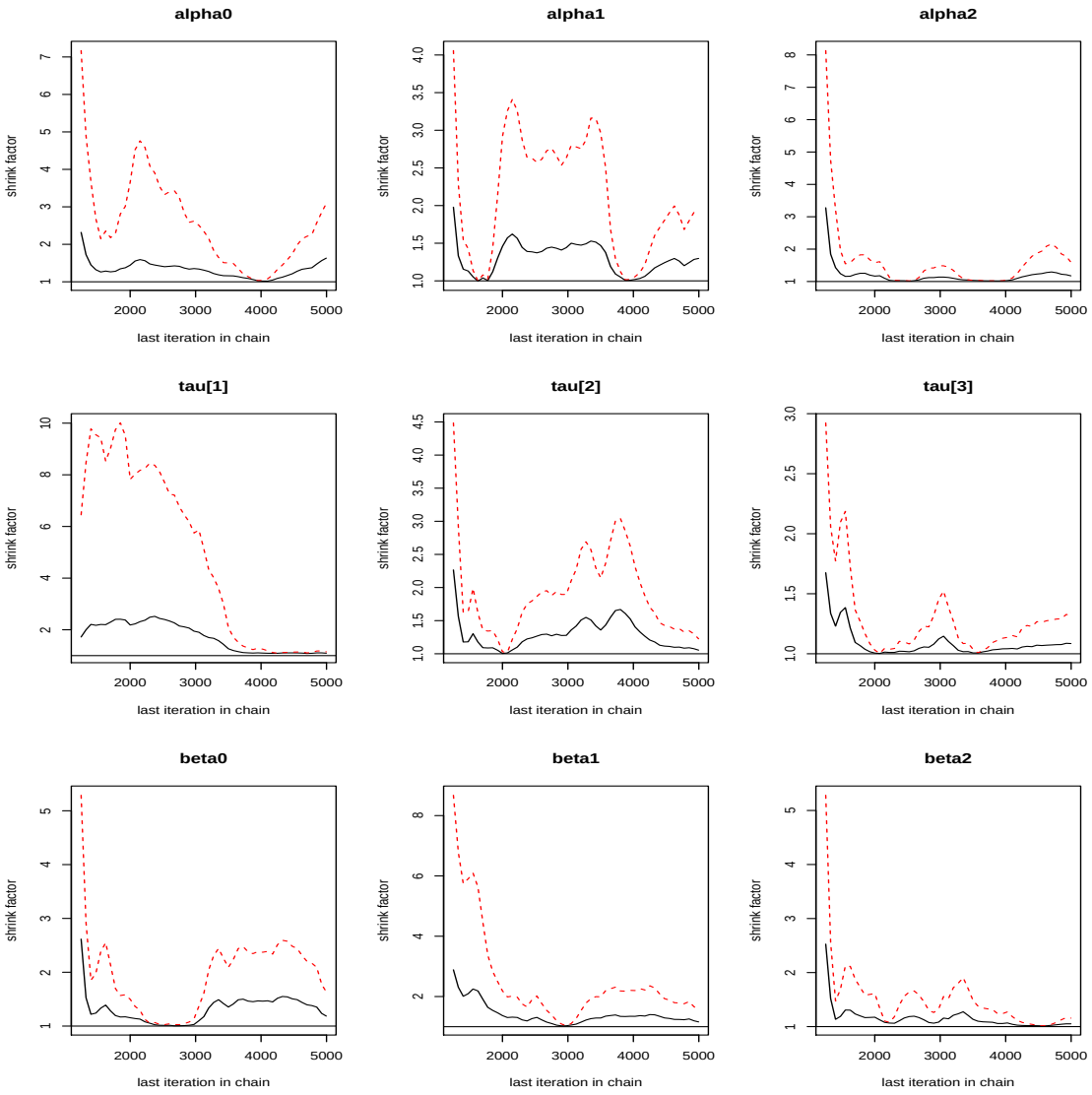


Figure 6.11: Gelman plots showing the PSRF values across iterations in the mixture model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the TTR. Plotted on the first row are the coefficients of growth (alpha1) and log-RD (alpha2) on π, on the second row are the factor loadings, while the third row shows Gelman plots for the coefficient of growth (beta1) and log-RD (beta2) in the TTR model.

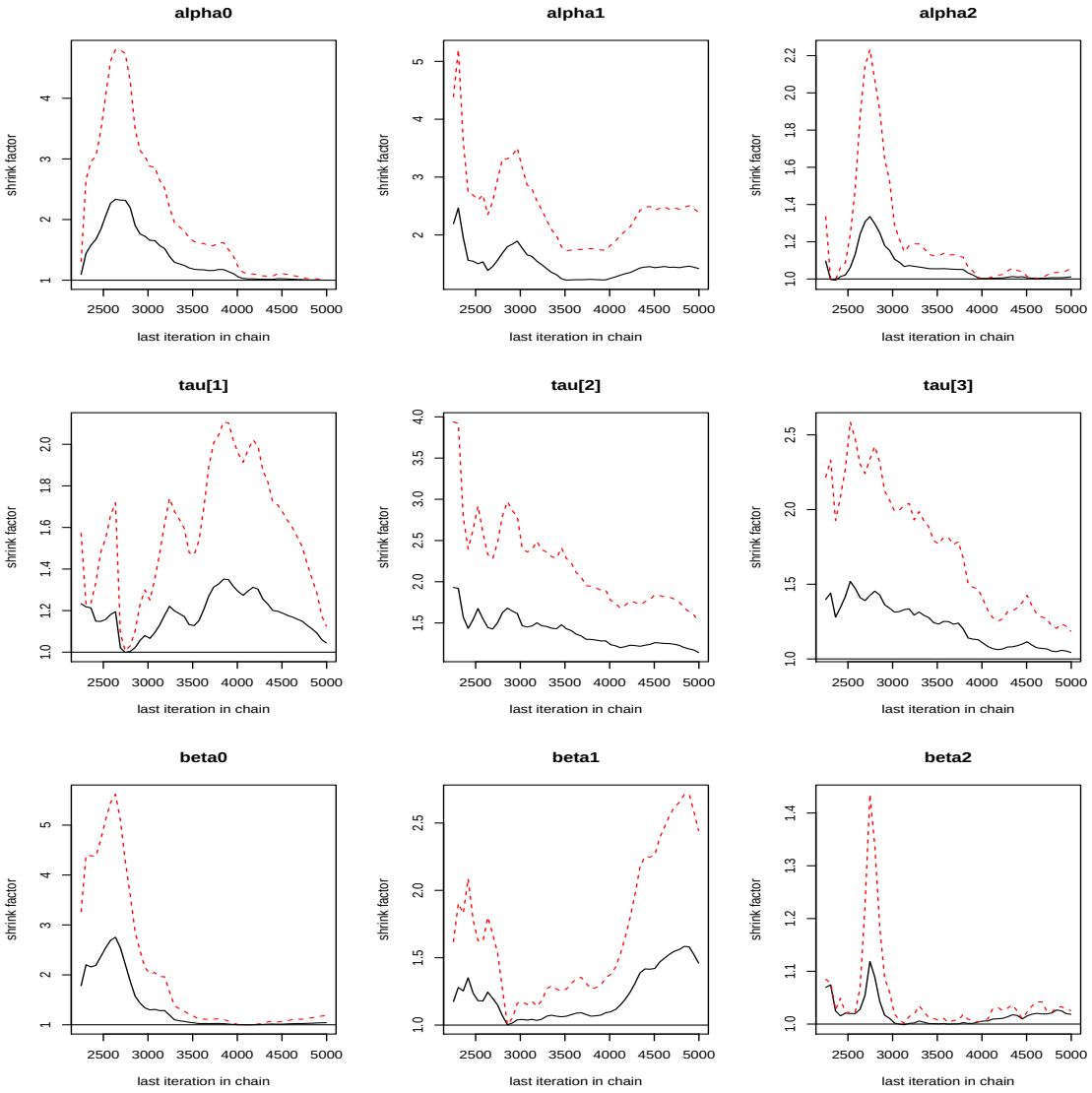


Figure 6.12: Gelman plots showing the PSRF values across iterations in the PTC model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the TTR. Plotted on the first row are the coefficients of growth (alpha1) and log-RD (alpha2) on π, on the second row are the factor loadings, while the third row shows Gelman plots for the coefficient of growth (beta1) and log-RD (beta2) in the TTR model.

fitted using the code in Appendix D.2. The idea in these models was again to investigate the role of the growth rate as a fixed covariate in the cure rate models similar to Section

4.8.

In both the mixture and PTC models, the probability of attaining OS similar to disease-free individuals in the general population was lower for those with faster growing tumours. The estimated log-OR in the population mixture model was -1.036 while it was -1.642 in the population PTC model. However, the CrIs in both model were wide and included a 0, suggesting that once again the possibility of a cure could not be established using these data. The posterior means of the coefficient of the growth rate also differed greatly between the two models, making it difficult to determine which of the two estimates was the true effect size. On the other hand, the growth rate was predictive of OS in the population mixture model log-HR 1.270 (0.164, 3.230), but not in the PTC model, log-HR 0.496 (-1.288, 2.047) where the effect of the growth rate on OS was less than half that in the mixture model, with a CrI that supported the null by including a zero. The estimate in the mixture model was also bigger than the standard Weibull model estimated log-HR of 0.844 (0.202, 1.811) in Table 6.3. The wider CrI in the mixture model points to issues with model convergence and we therefore recommend treating the effect size in this model with caution. It can of course still be surmised that the risk of death among those not cured was high, driven mainly by the tumour growth rate.

The estimates in the measurement parts for both models showed an association of the growth rate and the observed variables, with similar estimates to those from the earlier models except for log-paraprotein which was not associated with the growth rate in the mixture model as the CrI included a zero. This shows that the population mixture model struggled to identify the underlying latent growth rate through the 3 observed variables. In terms of DIC, there was little separating the mixture and PTC models. The posterior distributions of the growth rate in both models was roughly centred at 0, Figure 6.13. The scatter plots of π against the growth rate showed a decrease in the probability of a patient's OS returning to that of the general population if they had a fast

growing tumour as expected. Again the distribution of the growth in both models did not span the whole range of the assumed $N(0,1)$ distribution, showing that while the models hinted at the existence of the latent growth rate, the available could not fully support the proposed model.

For the mixture model, the Gelman plots for the parameters hinted at good convergence in general. The plot for the factor loading for paraprotein (tau[1]), the intercept in the model for π (alpha0) and the estimate of the log-OR for growth in the model for π (alpha1) depicted convergence problems alongside the intercepts in the measurement models for the log-transformed paraprotein and beta2, Figure 6.14. However, these issues could be remedied by running longer MCMC chains. In the PTC model, problems with convergence were evident in the Gelman plots of the factor loadings on both paraprotein and beta2 as well as the intercept in the model for π. The PSRFs for these parameters were clearly above the threshold of 1.1. In the same vein, the plots for the intercepts in the measurement models showed wriggling of the PSRF across iterations, Figure 6.15. This once again re-affirms the superiority of the mixture model over the PTC in estimating the proportion whose OS returns to that of disease-free individuals in the general population.

6.10 Tumour growth rate and log-RD as covariates in population mixture and PTC models

The role of growth rate adjusted for log-RD on the proportion whose OS returns to that of the general population was then investigated using the Bayesian Weibull mixture and PTC models, Table 6.8. The models were fitted using the code in Appendix D.2.

In the model for π, a SD increase in the growth rate was associated with a decrease in the probability of attaining OS similar to disease-free individuals in the general population; log-OR -0.759 and -1.216 in the population mixture and PTC models respectively.

	Weibull population mixture	**Weibull population PTC**
Parameter	Estimate (95 % CrI)	Estimate (95 % CrI)
Cure model (π)		
Intercept (α_0)	-3.408 (-9.132, -0.385)	-2.241 (-5.646, 0.045)
Growth (α_1)	-1.036 (-6.126, 3.451)	-1.642 (-5.457, 2.081)
OS model (λ)		
Intercept (β_0)	-3.659 (-5.995, -2.554)	-4.427 (-6.077, -2.844)
Growth (β_1)	1.270 (0.164, 3.230)	0.496 (-1.288, 2.047)
Shape (γ)	1.826 (1.258, 2.924)	1.822 (1.391, 2.398)
SEM part		
Paraprotein		
Intercept (μ_1)	0.939 (0.800, 1.075)	0.941 (0.795, 1.086)
Growth (τ_1)	0.220 (-0.071, 0.449)	0.340 (0.044, 0.664)
Beta2		
Intercept (μ_2)	1.584 (1.523, 1.644)	1.584 (1.527, 1.637)
Growth (τ_2)	0.151 (0.037, 0.325)	0.137 (0.003, 0.285)
Albumin		
Intercept (μ_3)	1.489 (1.464, 1.511)	1.489 (1.466, 1.509)
Growth (τ_3)	-0.038 (-0.066, -0.011)	-0.041 (-0.075, -0.009)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.244 (1.021, 1.499)	1.179 (0.846, 1.454)
Var(Beta) (ϵ_2)	0.233 (0.165, 0.294)	0.238 (0.183, 0.293)
Var(Albumin) (ϵ_3)	0.029 (0.023, 0.034)	0.028 (0.023, 0.034)
DIC	7,735	7,738

Table 6.7: Bayesian Weibull mixture and PTC models for the proportion with OS similar to general population fitted to Myeloma data with tumour growth rate as covariate on both π and the OS fitted to the Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-OR as a coefficient for growth (α_1) are presented. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

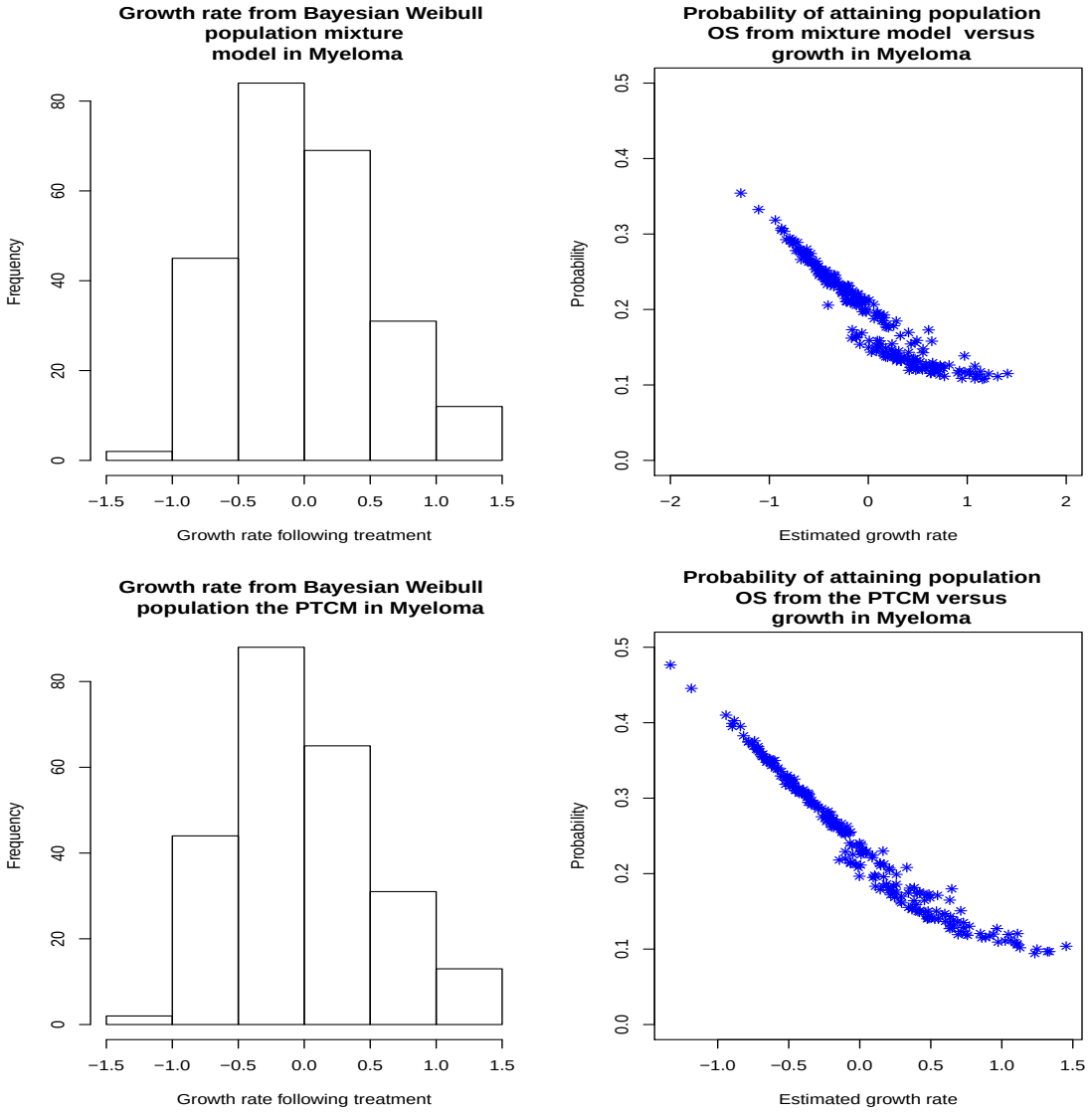


Figure 6.13: Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of attaining population OS versus the growth rate from Weibull population mixture and PTC models in the Myeloma dataset

However, the CrIs for both estimates were wide and included a zero. With respective to log-RD, a higher disease burden was associated with an increased probability of attaining the population level OS when growth was taken into account; log-OR 0.693 (-1.705, 5.5138) in the mixture model and 0.582 (-1.622, 4.535) in the PTC model.

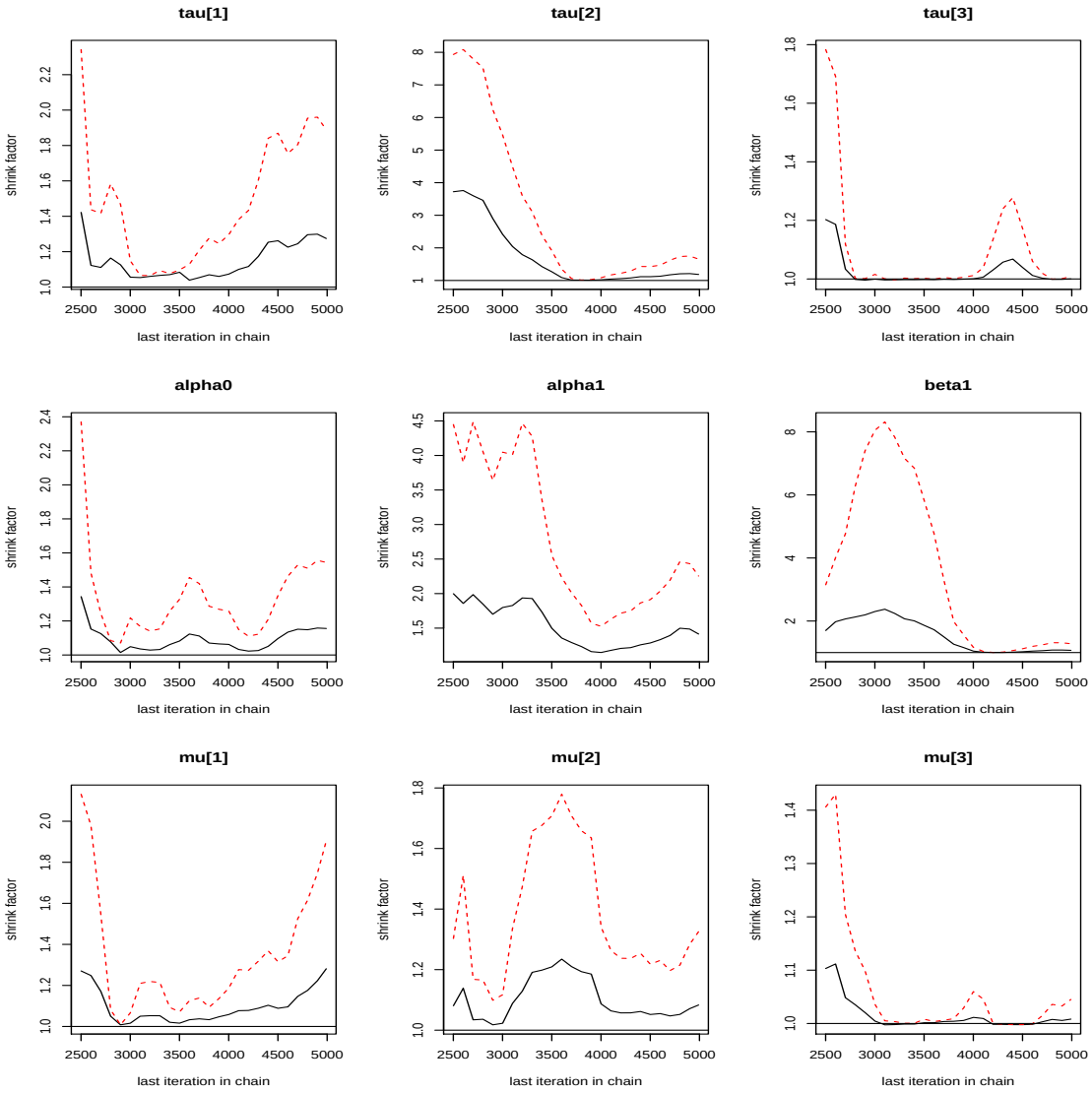


Figure 6.14: Gelman plots showing the PSRF values across iterations in the population mixture model for the cured proportion in the Myeloma dataset with growth as a covariate on both π and the OS distribution. Plotted on the first row are the plots for the factor loadings, while coefficients of growth (alpha1) on π, and (beta1) on the OS are shown on the second row. Finally, the plots for the intercepts from the measurement models are on the third row.

These estimates imply having a larger disease burden after treatment was protective, contrary to the running hypothesis in this thesis. These estimates could have resulted

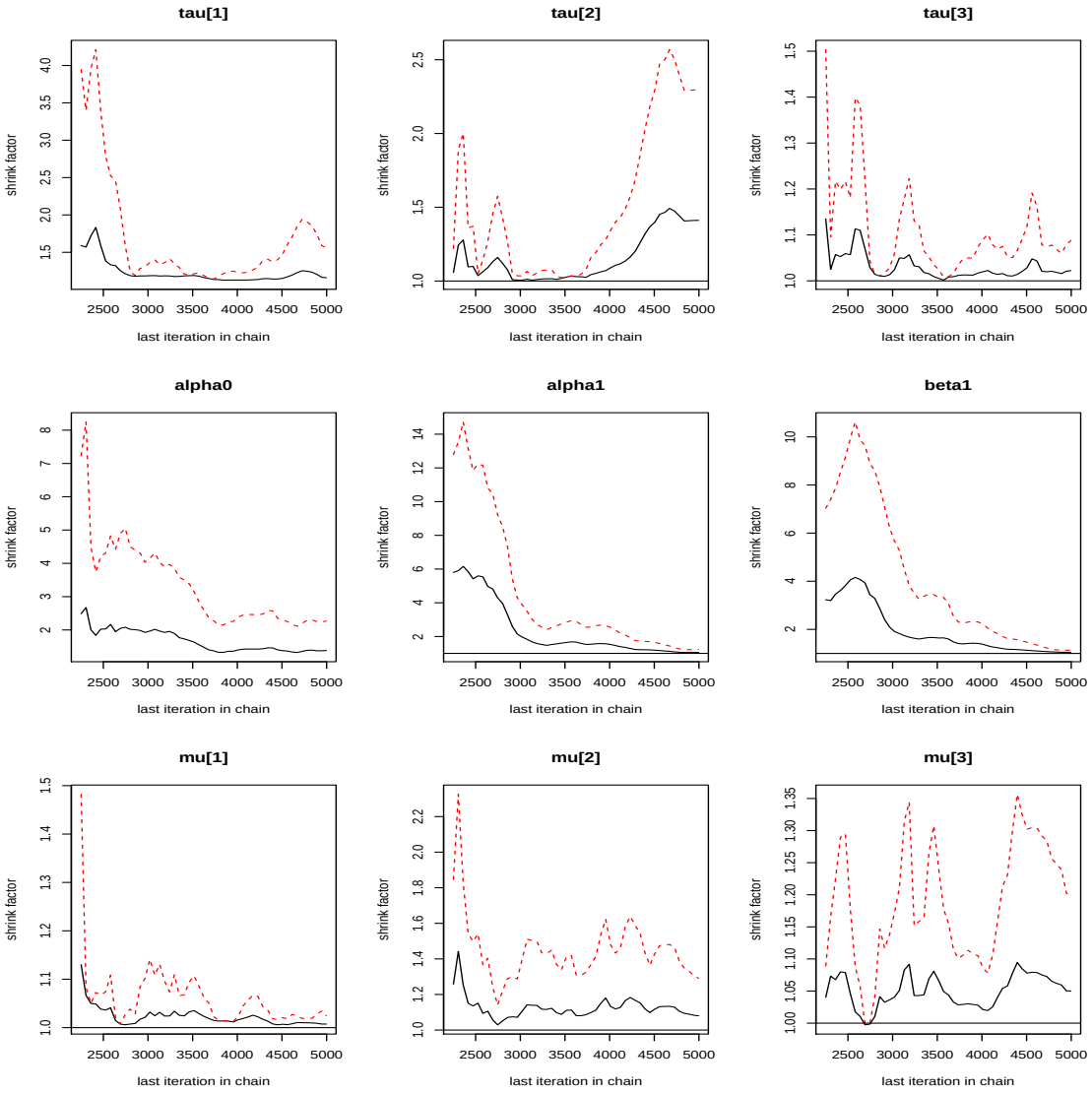


Figure 6.15: Gelman plots showing the PSRF values across iterations in the population PTC model for the cured proportion in the Myeloma dataset with growth as a covariate on both π and the OS. Plotted on the first row are the plots for the factor loadings, while coefficients of growth (alpha1) on π, and (beta1) on the OS are shown on the second row. Finally, the plots for the intercepts from the measurement models are on the third row.

from problems with model fitting or the data not supporting the existence of a proportion

of patients whose OS would return to that of age-sex matched individuals in the general

population.

In the OS models, an SD increase in the growth rate was predictive a lower risk of death in the mixture model, log-HR -0.198, and a higher risk of death in the PTC model, log-HR 0.566 when log-RD was taken into account. The PTC model estimate was similar to the effect in the standard OS model of Section 6.6, while the mixture model estimate could be a manifestation of model fit problems. An increase in the log-RD after treatment was associated with a higher risk of death, log-HRs 0.223 and 0.427 in the mixture and PTC models for OS respectively. Adjusting for the growth tended to reduce the effect size of the log-RD on the log-hazard of dying as expected.

Despite the obvious problems with convergence, the expected inverse relationship between the growth and π was visible from the scatter plots, Figure 6.16. In these plots, the probabilities were not uniformly decreasing with increasing growth rate especially in the mixture model which had individuals with tumour growth rates more than 1 SD above the 0 having a high probability of attaining the population level OS. In the PTC, the distribution of the growth rate ranged from -1 to 1.5 SDs of 0 which was another manifestation of the data not supporting the proposed model.

From the SEM part, it was clear that the underlying latent variable representing growth was not fully manifested by the three variables paraprotein, beta2 and albumin. In all measurement models, the CrI for the estimated coefficient of growth on each of the log-transformed paraprotein, beta2 and albumin included a 0, suggesting no association between the growth and each of these variables. This could have contributed to the problems with model convergence due to the relatively small dataset from which we aimed to estimate parameters using these complicated models. Gelman plots showing PSRFs for some select parameters across iterations in the MCMC showed that there were convergence problems for these models. In the mixture model, the plots for all parameters except the intercept in the models for π (alpha1) and the OS (beta1) showed PSRFs well above 1, Figure 6.17. This means the MCMC chains of the conditional

posterior distributions for these parameters did not tend towards the target as they continued to explore different areas of the posterior. From the PTC model, the Gelman plots depicted a similar trend, with PSRFs way above 1 especially for the factor loadings for paraprotein and beta2, as well as for the coefficient of log-RD in both the model for π and OS, Figure 6.18.

Both models struggled to estimate the proportion whose OS returns to that of disease-free individuals in the general population. In terms of the DIC, the PTC model performed better after 5,000 iterations when compared to the mixture model. Looking at the SEM part, estimated of the coefficients of growth in the measurement equations, especially in the mixture models, were clearly different from those in the earlier models, highlighting the fit problems further.

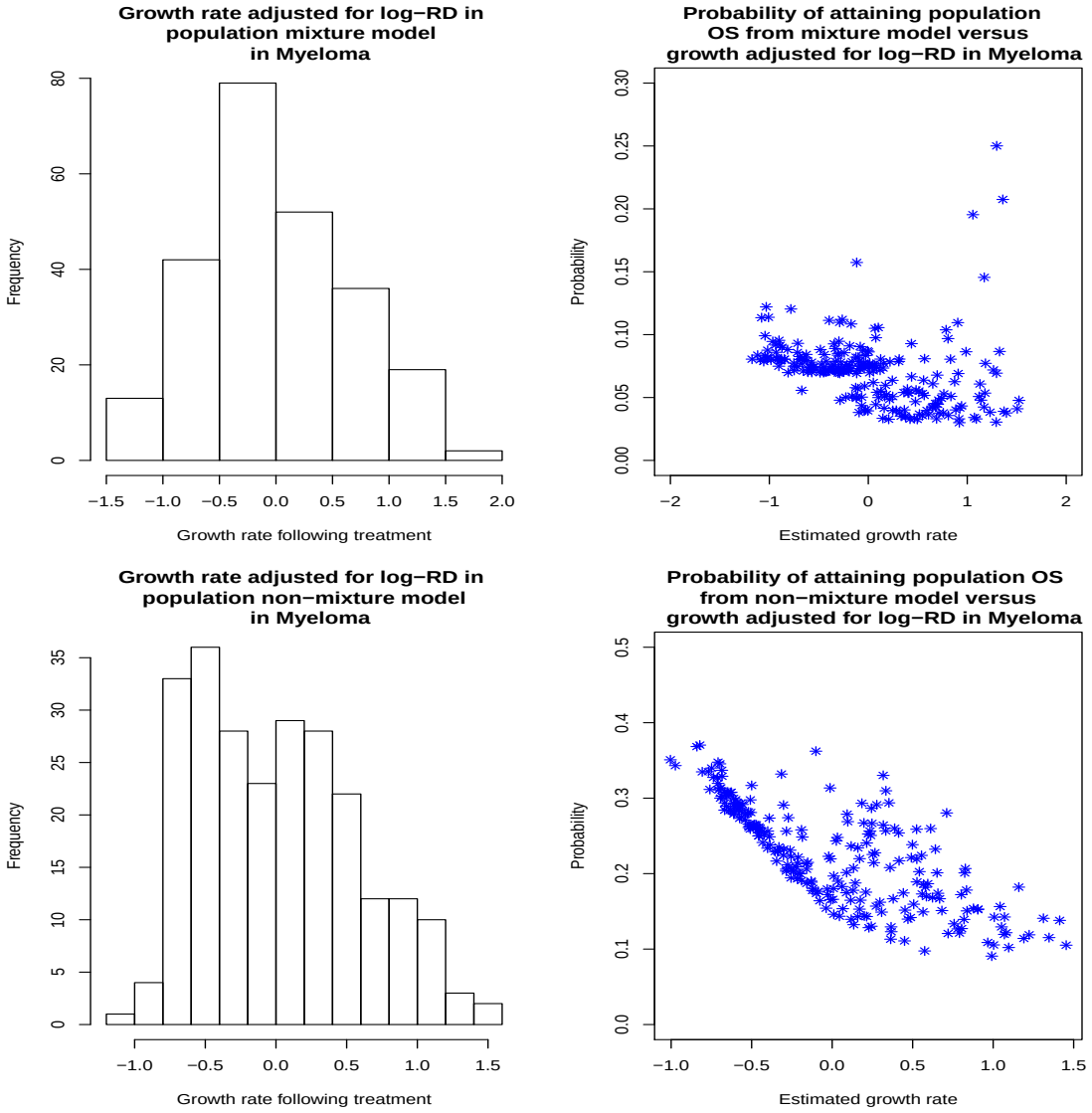


Figure 6.16: Histograms depicting the posterior distribution of the growth rate as well as scatter plots of the probability of attaining population OS versus the growth rate adjusted for log-RD from Weibull population mixture and PTC models in the Myeloma dataset

6.11 Checking for the possibility of a cured proportion

The cure rate models enabled us to investigate the role of growth, through SEM, and log-RD on the probability of not relapsing again or dying from Myeloma as well mod-

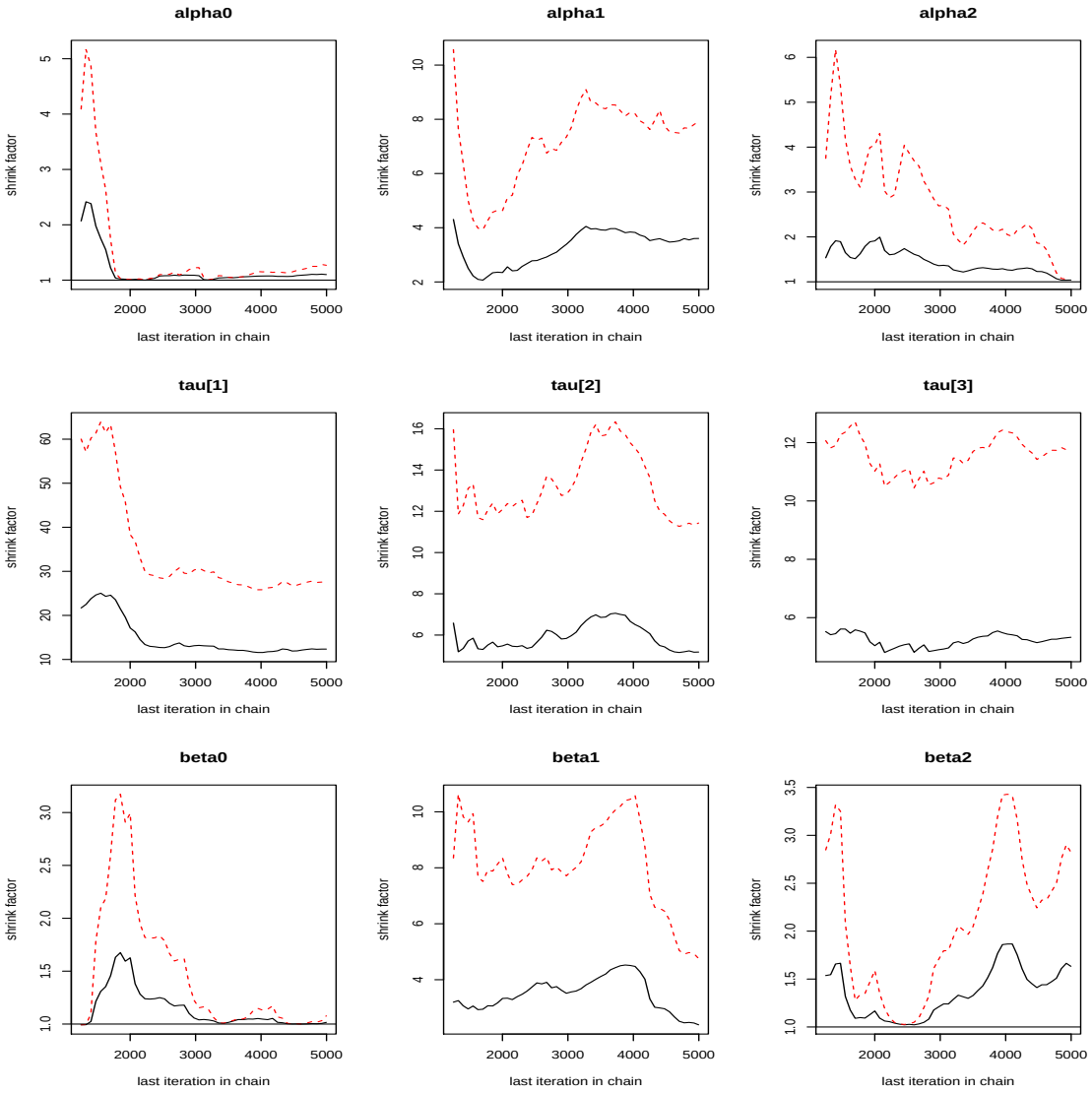


Figure 6.17: Gelman plots showing the PSRF values across iterations in the population mixture model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the OS. Plotted on the first row are the intercept (alpha0), the coefficient of growth (alpha1) and log-RD (alpha2) on π, plots of the factor loadings tau[1], tau[2] and tau[2] are on the second row while the third row has plots for the intercept beta0, the coefficient for growth (beta1) and log-RD (beta2) in the OS model.

elling the TTR and OS for those who were not cured especially through the mixture models. While the models provided reasonable estimates of the parameters in some in-

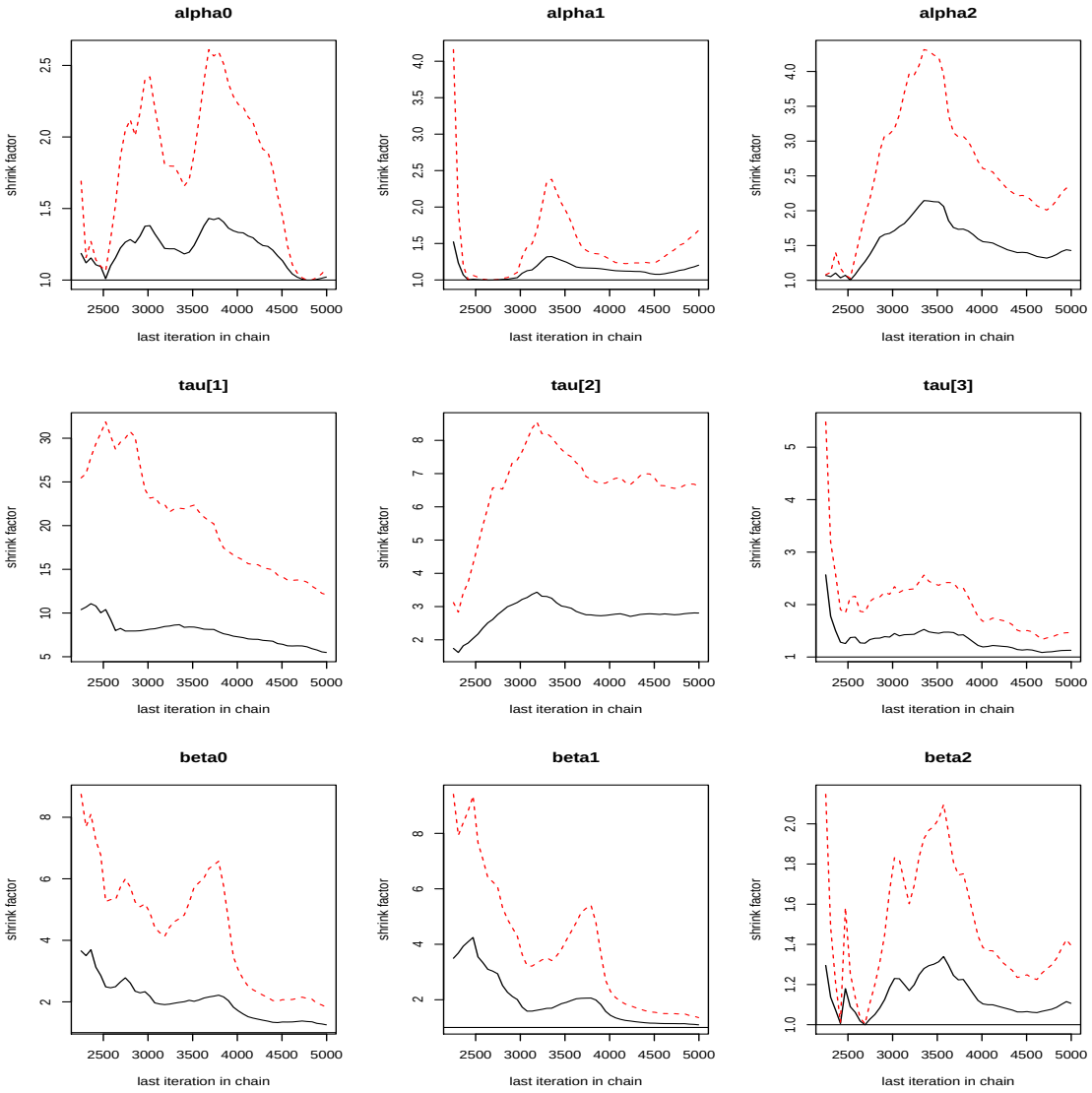


Figure 6.18: Gelman plots showing the PSRF values across iterations in the population PTC model for the cured proportion in the Myeloma dataset with growth and log-RD as covariates on both π and the OS. Plotted on the first row are the intercept (alpha0), the coefficient of growth (alpha1) and log-RD (alpha2) on π, plots of the factor loadings tau[1], tau[2] and tau[2] are on the second row while the third row has plots for the intercept beta0, the coefficient for growth (beta1) and log-RD (beta2) in the OS model.

stances, the nature of the available data itself needed to be checked to confirm whether cure was possible.

We used trace and density plots from the Bayesian Weibull mixture model for TTR and the Bayesian Weibull population mixture model for OS to determine this, Figure 6.19. The trace plots for π and λ in both models suggested that the MCMC chains converged to the posterior. The estimated proportion who would never relapse and those whose OS would resemble that of age-sex matched individuals around a third in this subset of patients. However, from the posterior density plots, it was clear that the distributions of π did not centre around the posterior mean in each model. The CrIs for π in the mixture model applied to TTR data spanned the whole range of values between 0 and 100%, while in the model applied to OS data, the CrI ranged from 0 - 50%. The estimates of π in these models were respectively above the 20% proportion who had not relapsed in the model applied to TTR and below around 50% who were still alive at the end of follow-up as shown in Figure 6.1 in Section 6.1.1. This highlights therefore, that the data did not support the existence of a proportion who might have been cured following treatment looking at either outcome, which might have led to the various problems with model encountered when using cure rate models in the previous sections.

6.12 Relaxing the multivariate normal assumption in SEM

Our approach so far has focused on modelling the growth rate from continuous observed variables assumed to have a multivariate normal distribution in the set-up described in Section 6.1.1. It was apparent from the plotted distributions of the observed variables assumed to depict tumour growth in Figure 6.2, that paraprotein had a bimodal distribution because 45% (110/243) of the patients included in this analysis did not have detectable paraprotein in their blood at baseline and were therefore assigned a value of 0 for this variable.

Several methods for modelling data that has many zeroes have been investigated in-

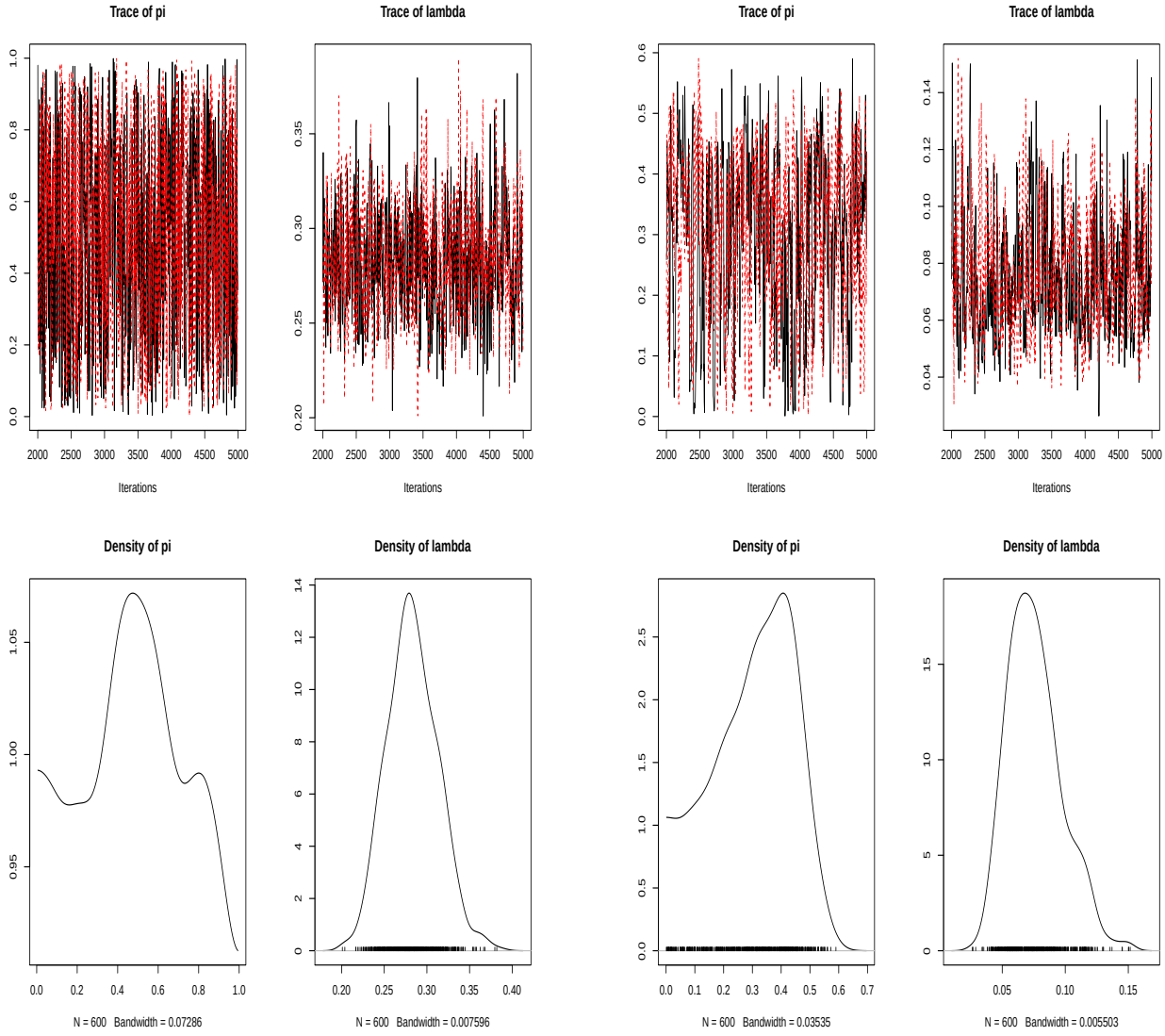


Figure 6.19: Trace and density plots of the scale (lambda) and π (pi) from the Bayesian Weibull mixture model for TTR (A) and Bayesian Weibull population mixture model for OS (B) in the Myeloma dataset.

cluding calibration methods that aim to take into account situations where the normality assumptions in linear models might be violated [154]. In SEM, methods exist for modelling latent variables using a combination of continuous, categorical and even ordinal observed variables [142]. Taking advantage of this feature, we can model paraprotein as a dichotomous random variable in a SEM.

We briefly investigate how the problem with many zeroes in paraprotein may be handled in a SEM model for TTR and OS in Myeloma using (5.3) in the model relating paraprotein as a binary observed variable (0 if no paraprotein, or 1 if greater than 0) and the tumour growth rate and at the same time use (5.1) to relate the continuous observed variables beta2 and albumin to the growth rate as before. The models with this mixture of a one binary (paraprotein) and two continuous (beta2 and albumin) observed variables were fitted in Stata using **gsem** which allows for a mixture of binary and continuous observed variables. Because paraprotein was a binary variable, there was no estimate of its variance in the measurement model (ϵ_1) in these models. Again, a Weibull distribution was assumed for the TTR and OS. Further, we fit both univariable, with growth as the only covariate on the TTR and OS, as well as multivariable models where both the effects of growth and log-RD on these outcomes were estimated.

Estimates from these models included log-HRs relating the growth rate and/or log-RD to the TTR and OS as well as the shape parameter (γ) in the Weibull model. The model formulation in the SEM part in this case included the usual linear regression models for the relationship between beta2 and albumin and the growth rate, and a logistic model relating the binary paraprotein and growth.

6.12.1 The effect of growth and log-RD on TTR with paraprotein as binary covariate

We again fit the TTR model assuming a Weibull distribution with growth as a covariate, as well as both growth and log-RD in a multivariable model with the latent growth rate

modelled through the observed variables as discussed. Estimates from these models are shown in Table 6.9.

In the model with growth only, the estimates in the TTR model were similar to those obtained from the original model where paraprotein was on the continuous scale. Each SD increase in the growth rate was associated with a significant 1.304 increase in the log-hazard of relapsing, a slight decrease from the log-HR of 1.385 in the original model assuming all markers were continuous in Section 6.3. The growth rate was also associated with each of the observed variables via the measurement equations. For instance, the probability of not having the marker paraprotein in the blood following treatment decreased with each SD increase in the growth rate, log-OR -0.591 (-0.992, -0.189). In the linear models, a faster growth rate was associated with higher beta2 and lower albumin levels similar to the trend seen in the original model. The variance estimates were again similar to those in the original model. We were thus able to model the effect of the latent growth rate on its own, on the TTR while at the same time showing strong associations between the growth rate and its 3 observed markers.

Including log-RD alongside the growth as covariates on the TTR resulted in a slight reduction in the effect of the growth on the TTR, estimated log-HR 1.086. In this model, each unit increase in the log-RD was associated with a 0.329 (-0.065, 0.722) increase in the log-hazard of relapsing. This effect was bigger than that in the original model (0.280) where paraprotein was a continuous variable, but it was not significant in either model. All the other estimates in the SEM part were similar to those in the original model, suggesting that the two approaches support the existence of the latent construct representing growth which is itself a associated with the TTR in Myeloma through the continuous observed variables only, or through a combination of binary and continuous markers of the growth rate.

With the added complexity by including both the growth rate and log-RD in the TTR model, interest might be on checking whether the simpler model was better than the

more complicated model with both covariates using the likelihood ratio test. The difference in the LL values between the two models was 1.4, which when doubled to account for the additional parameter in the multivariable model was 2.8 which was less than the ($\chi^2_{(1)} = 3.84$). From a purely statistical point of view, it would suffice to model the TTR using growth as the only covariate by choosing the simpler model. However, both growth and log-RD are important predictors of outcomes in cancer. We would thus recommend characterising the TTR using both covariates taking advantage of these methods.

6.12.2 The effect of growth and log-RD on OS with paraprotein as binary covariate

The model was finally used to investigate the association between tumour growth and OS, Table 6.10. The effect size of the growth rate was almost the same as that in the original model in Table 6.3. Each SD increase in the growth rate was associated with a significant 0.525 increase in the log-hazard of dying in this model, an effect almost the same as that reported in the original model where paraprotein was a continuous variable (log-HR, 0.527). Again in the logistic model relating the growth rate to paraprotein as a binary covariate, each SD increase in the growth rate was predictive of a lower probability of having zero levels of paraprotein in the blood after treatment. However, this effect was not significant. In the SEM part, an increase in the growth rate was predictive of lower chances of not having paraprotein after treatment, higher beta2, and lower albumin levels. The association of the growth rate and paraprotein in the logistic model was also not significant.

When log-RD was included in the model, an increase in the growth rate was associated with a slightly bigger increase in the hazard of dying when log-RD was taken into account. Compared to the model based on continuous biomarkers only, the log-HR for paraprotein was slightly bigger 0.306 in the new model versus 0.292 in the origi-

nal model. The same associations of the growth rate and the observed variables was seen in the SEM part, with a non-significant log-OR in the model relating the growth rate to paraprotein as a binary variable. Again looking at the LL values, there was little separating the two models and we would recommend using both growth and log-RD in modelling the OS in Myeloma.

The variances beta2 and albumin were similar in both models. The SEM part in these models hinted at the same underlying latent construct representing the growth rate which was itself a predictor of the OS alongside log-RD. However, the inclusion of a 0 in the 95% CI in the model relating the binary paraprotein and the growth rate was consistent with what was seen in the other OS models where the data did not fully support the hypothesised latent growth rate observed through the 3 markers.

6.12.3 Comparison with SEMs based on continuous variables

To compare the models where paraprotein is a binary covariate with those where paraprotein was a continuous variable, we calculated the AIC using (2.35) from the LL values and number of parameters for each of the TTR and OS models.

The AIC in the TTR model with growth as the only covariate and paraprotein as a continuous variable was calculated AIC as 1,651.2, while that from the same model but with paraprotein as a binary covariate was 1,236. In the TTR model with both growth and log-RD as covariates, the AIC was 1,651.4 when paraprotein was included as a continuous observed variable, and 1,235.4 when it was binary. In the OS models, with growth as the only covariate and paraprotein as a continuous variable, the AIC was 1,486.2, while when paraprotein was binary, the AIC was 1,067.8. Finally, in the model for OS with both growth and log-RD with paraprotein as a continuous covariate, the AIC was 1,486.4 while with paraprotein as a binary variable, the AIC was 1,067.8.

In all the OS and TTR models, the change in AIC between models with growth only and with growth and log-RD as covariates was small regardless of whether paraprotein

was a continuous or binary variable. This re-affirms our belief that both growth and log-RD are important predictors of these outcomes. The AICs in the models where paraprotein was a binary variable were less than that in the models where paraprotein was continuous. Moreover, these models had one less parameter to be estimated than those where paraprotein was a continuous variable. All things being equal, we would choose the less complicated models with paraprotein as a binary variable based on the smaller AIC values over those where paraprotein is continuous. This aspect of the work could be investigated further and is beyond the scope of this thesis, as is the use of SEMs with both continuous and binary observed variables to model the effect of growth in cure rate models.

6.13 Summary

In this chapter, we applied SEMs to the Myeloma data, focusing on the role of the tumour growth modelled via the observed variables log-paraprotein, log-beta2 and log-albumin on the various TTE outcomes. This approach provides a departure from the usual way of analysing survival outcomes in RCTs involving cancer patients, and is motivated by opportunities provided by a combination of SEM and survival methods, and the presence of biomarkers in the Myeloma dataset that are can be said to represent tumour growth. SEM methods have been successfully used in medical research, including in studies for rheumatoid arthritis [155], where latent variables such as physical disability have been investigated based on a number of observed variables.

In the analyses from Chapters 3 to 4, we investigated the role of log-RD and the three variables which are markers of tumour growth on the various outcomes, mostly as independent covariates. The SEM approach used here provided a unique opportunity to explore how the three variables co-vary to bring out the underlying latent variable representing growth. When modelled together with log-RD in TTR, OS or cure rate

models, the latent variable growth always appeared to have a considerable effect on the outcomes, confirming that the growth rate of the tumour following treatment is an important predictor of survival. While this growth rate is not usually observed in practice, we are able to model its effect on TTE outcomes using SEMs by exploring the covariance structure between related variables that we believe are related to growth. Rather than adjusting the effect of log-RD for each of paraprotein, beta2 and albumin, we exploited the contribution of each of these variables to the growth which we in-turn treated as a fixed covariate in the TTE model. The SEM approach used in this thesis reiterates that using the biomarkers to represent the growth rate results in a stronger association with the outcomes than when each of these markers is used singly or together in multivariable models [57].

While the novel methods provide the unique opportunity to model the effect of tumour growth on the outcomes, there were challenges with the available data which made it difficult for us to fully exploit the strength of these methods. The biggest issue was that of missing data: the final dataset used in these analyses had data from 243/427 (56.9%), missing out on 43% of the full dataset. We thus fit the models to smaller datasets which could have led to problems with obtaining precise estimates. While the translocations would have provided good markers of the tumour growth rate, the amount of missing information for these variables made it impossible to use them in these models. Unlike in the models fitted in Chapters 3 and 4, there are no standard methods for handling missing data in the standard statistical software. We would still be able to use MI to impute a number of datasets, fit the models to each of them and then combine the output using Rubin's rules to get around this problem. Alternatively, Bayesian methods for handling missing data in SEMs, which have already been investigated, would be used. Both of these approaches were not the focus of this thesis but could be investigated in future alongside SEMs that combine binary, continuous and categorical observed variables. The second issue to bear in mind was to do with

the relatively weak correlations between the three observed variables themselves that together manifest the growth rate. The correlation was weakest between paraprotein and beta2 (-0.02), while the correlation between paraprotein and albumin was weak (-0.12), and it was moderate between beta2 and albumin (-0.26). This was something worth bearing in mind when interpreting results from these models. On top of this, log-transforming the three observed variables to satisfy the multivarite normality assumption did not achieve the desired effect for paraprotein where close to half of the patients did not have detectable levels in their blood. Implementing SEMs with paraprotein as a binary covariate went some way in dealing with this issue. Despite these problems, the models showed that the latent growth modelled through the three observed variables had an effect on both TTR and OS outcomes. The final potential pitfall relates to fitting cure rate models to the data. The number of events in terms of those who did not relapse was 56 (23%) when looking at TTR, and the number who died over the follow-up period was 94 (39%). It was thus not possible to estimate the proportion who are cured following treatment when a few of the patients using the complicated models while allowing for the latent variable which was evident in the analysis involving cure rate models. With OS as the outcome, there were challenges with fitting cure rate models to the bigger dataset in Chapter 4, which was also evident in this smaller dataset.

Due to the complexity of the models, there were some challenges in fitting some of the models especially those with OS. This might have made it especially hard to fit Cox PH and cure rate models with the tumour growth as a fixed covariate in MCMC as the counting process approach focuses on the event times themselves. On the other hand, where there were enough events as in the TTR, the models generally fitted well and it was possible to assess the possible role of both the growth rate and the log-RD on the various TTE outcomes. We also demonstrated that we can use a mixture of SEMs with binary and continuous observed variables to model the effect of tumour growth on the

TTR and OS in Myeloma using the available software. Extensions of these methods to cure models can also be made.

In this chapter, we have used well established methods that combine SEMs and TTE methods alongside Bayesian equivalents of the methods to model the TTR, OS as well as the possibility of a cure using cure rate models. The Weibull ML model fitted in Stata provides a standard way of analysing TTR or OS where we believe there is a latent variable such as growth in our case, as long as there are biomarkers available. In the Stata software the available routines only allow for SEMs with TTE outcomes assumed to come from parametric distributions such as the Exponential and Weibull. In this thesis, we showed that we can also combine SEMs and the Cox PH model to model both the TTR and OS in a Bayesian setting. We have also demonstrated that we can further extend the SEMs to allow for the modelling of latent variables as covariates in cure rate models in a Bayesian setting. This work therefore sets the stage where these methods can be fully exploited in larger datasets where there are underlying latent variables known to influence survival outcomes in RCTs.

	Weibull population mixture	**Weibull population PTC**
Parameter	Estimate (95 % CrI)	Estimate (95 % CrI)
Cure model (π)		
Intercept (α_0)	-3.066 (-7.167, -0.248)	-2.124 (-5.958, -0.220)
Growth (α_1)	-0.759 (-6.014, 5.007)	-1.216 (-4.950, 2.453)
Log-RD (α_2)	0.693 (-1.705, 5.138)	0.582 (-1.622, 4.535)
OS model (λ)		
Intercept (β_0)	-2.805 (-3.597, -2.097)	-4.416 (-7.162, -2.876)
Growth (β_1)	-0.198 (-1.179, 0.720)	0.566 (-0.915, 2.324)
Log-RD (β_2)	0.223 (-0.330, 0.772)	0.427 (-0.600, 1.420)
Shape ((γ))	1.530 (1.223, 1.951)	1.909 (1.344, 3.007)
SEM part		
Paraprotein		
Intercept (μ_1)	0.925 (0.803, 1.052)	0.920 (0.784, 1.077)
Growth (τ_1)	0.337 (-0.167, 0.823)	0.557 (-0.008, 1.025)
Beta2		
Intercept (μ_2)	1.585 (1.522, 1.645)	1.581 (1.511, 1.649)
Growth (τ_2)	-0.072 (-0.341, 0.195)	0.096 (-0.072, 0.357)
Albumin		
Intercept (μ_3)	1.490 (1.469, 1.510)	1.492 (1.471, 1.513)
Growth (τ_3)	0.003 (-0.061, 0.065)	-0.029 (-0.062, 0.002)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	1.069 (0.573, 1.487)	0.903 (0.317, 1.461)
Var(Beta2) (ϵ_2)	0.231 (0.150, 0.300)	0.240 (0.145, 0.303)
Var(Albumin) (ϵ_3)	0.029 (0.024, 0.035)	0.029 (0.024, 0.036)
DIC	7,755	7,669

Table 6.8: Bayesian Weibull mixture and PTC models for the proportion with OS similar to the general population fitted to Myeloma data with tumour growth rate and log-RD as covariates on both π and the OS fitted to Myeloma data based on complete cases ($n = 243$). In the cure model, estimates of the intercept (α_0) and the log-ORs as coefficients for growth (α_1) and log-RD (α_2) respectively, are presented. In the OS model, the intercepts (β_0) and shape parameter (γ) from the Weibull models, and log-HRs, as coefficients for growth (β_1) and log-RD (β_2) respectively from all models, are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported

	(A) Growth only	**(B) Growth and log-RD**
Parameter	Estimate (95% CI)	Estimate (95% CI)
TTR model (λ)		
Intercept (β_0)	-1.709 (-2.399, -1.019)	-1.392 (-1.976, -0.809)
Growth (β_1)	1.304 (0.067, 2.541)	1.086 (0.167, 2.006)
Log-RD (β_2)	-	0.329 (-0.065, 0.722)
Shape (γ)	1.622 (0.977, 2.694)	1.490 (1.003, 2.215)
SEM part		
Paraprotein		
Intercept (μ_1)	-0.204 (-0.478, 0.069)	-0.201 (-0.470, 0.068)
Growth (τ_1)	-0.591 (-0.992, -0.189)	-0.516 (-0.937, -0.094)
Beta2		
Intercept (μ_2)	1.586 (1.523, 1.649)	1.586 (1.523, 1.649)
Growth (τ_2)	0.128 (0.022, 0.234)	0.148 (0.032, 0.264)
Albumin		
Intercept (μ_2)	1.489 (1.471, 1.508)	1.489 (1.471, 1.508)
Growth (τ_3)	-0.054 (-0.087, -0.021)	-0.060 (-0.097, -0.023)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	-	-
Var(Beta2) (ϵ_2)	0.235 (0.192, 0.288)	0.230 (0.184, 0.286)
Var(Albumin) (ϵ_3)	0.019 (0.015, 0.024)	0.018 (0.014, 0.024)
LL	-607.0	-605.7

Table 6.9: Weibull ML models for the TTR fitted to Myeloma data based on complete cases ($n = 243$) with tumour growth rate only (A) and tumour growth rate and log-RD (B) as covariates on the TTR with paraprotein treated as binary. In the TTR model, the intercept (β_0) and shape parameter (γ), and the log-HR, as a coefficient for growth (β_1) in (A) and for both growth (β_1) and log-RD (β_2), are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

	(A) Growth only	**(B) Growth and log-RD**
Parameter	Estimate (95% CI)	Estimate (95% CI)
OS model (λ)		
Intercept (β_0)	-3.136 (-3.646, -2.626)	-2.944 (-3.506, -2.382)
Growth (α_1)	0.525 (0.126, 0.924)	0.528 (0.142, 0.914)
Log-RD (α_2)	-	0.306 (-0.120, 0.734)
Shape (γ)	1.455 (1.203, 1.758)	1.452 (1.202, 1.755)
SEM part		
Paraprotein		
Intercept (μ_1)	-0.192 (-0.449, 0.064)	-0.192 (-0.447, 0.064)
Growth (τ_1)	-0.235 (-0.669, 0.199)	-0.205 (-0.638, 0.228)
Beta2		
Intercept (μ_2)	1.586 (1.523, 1.649)	1.586 (1.523, 1.649)
Growth (τ_2)	0.246 (0.116, 0.377)	0.254 (0.116, 0.392)
Albumin		
Intercept (μ_3)	1.489 (1.471, 1.508)	1.489 (1.471, 1.508)
Growth (τ_3)	-0.075 (-0.113, -0.036)	-0.074 (-0.113, -0.035)
Variances		
$Var(\omega)$	1.0 (fixed)	1.0 (fixed)
Var(Paraprotein) (ϵ_1)	-	-
Var(Beta2) (ϵ_2)	0.191 (0.134, 0.271)	0.187 (0.127, 0.274)
Var(Albumin) (ϵ_3)	0.016 (0.011, 0.023)	0.016 (0.011, 0.024)
LL/DIC	-522.9	-521.9

Table 6.10: Weibull ML models for the OS fitted to Myeloma data based on complete cases ($n = 243$) with tumour growth rate only (A) and tumour growth rate and log-RD (B) as covariates on the OS with paraprotein treated as binary. In the OS model, the intercept (β_0) and shape parameter (γ), and the log-HR, as a coefficient for growth (β_1) in (A) and for both growth (β_1) and log-RD (β_2), are reported. For the SEM part, the intercept and coefficients from the measurement models are reported. Variances for each observed variable are also reported.

Chapter 7

Discussion

7.1 Introduction

This chapter summaries the work that has been done in this thesis and provides suggestions on how the proposed methods can be used in practice, while highlighting their strengths and limitations. Finally, further avenues for this research are proposed.

This thesis was focused on modelling the role of both observed and unobserved measures of the tumour namely the log-RD and tumour growth rate following treatment, on survival outcomes, with applications to Myeloma and CLL datasets.

Chapter 1 provided a brief background to the various approaches that are used to model important survival outcomes such as TTR, OS and where possible, the proportion of patients who are completely cured of the disease following treatment.

Using standard TTE methodology, the role of the log-RD on its own, and in conjunction with other covariates in predicting TTR and OS was investigated. In Myeloma, a higher disease burden, measured through the log-RD, had an adverse, though not statistically significant effect on TTR and OS. This is in agreement with what has been shown in the clinical literature that in Myeloma, the MRD is an important predictor of survival outcomes following treatment [156]. Those with advanced disease, detectable

paraprotein, and a higher beta2 had poorer outcomes in this group of patients while lower levels of albumin conferred a greater risk of relapsing or dying. Cytogenetic aberrations are important markers of how aggressively the tumour is growing and thus predictors of TTR and OS. They are particularly useful in classifying a patient's risk as low, intermediate and high risk of relapse or death [157]. However, in the Myeloma dataset, about 42-48% of patients did not have data on these cytogenetic markers although they were significantly associated with the outcomes in a CC analysis including only those patients who had data on these markers. In the CLL trials, comparisons made with respect to TTR and OS on sex, treatment given and trial under which the patient was recruited showed no association with either of the two outcomes. Those with higher log-RD after treatment had worse outcomes, consistent with what was seen in Myeloma. The effect of log-RD in CLL was significant for both the TTR and OS. Those with unmutated VH genes had poorer TTR and OS, consistent with what has been reported [158]. Patients with 17p deletion are known to have poor outcomes [159]. However, for this study, those with 17p deletion seemed to have better outcomes. This could have been due to the fact that only 12/310 did not have this trait thereby making it hard to observe heterogeneity with respect to this marker. To account for missing data, the final part of this chapter focused on fitting Bayesian models with imputation models for the missing covariates in both datasets. The resulting models showed that the effect sizes for paraprotein, beta2, albumin and the cytogenetic markers in Myeloma and VH mutation and 17p deletion in CLL were similar to those in the CC analysis, thereby implying the missing data was ignorable.

Chapter 4 was concerned with using cure rate models to estimate the proportion of patients who will remain relapse-free or those whose OS will match that of age-sex matched members of the general population following treatment. The proportions were estimated using mixture and PTC models - with both semi-parametric via a Cox PH model, and parametric Weibull models used to model the TTR for those not cured and

therefore bound to relapse and/or subsequently die following the treatment [18, 69, 26]. In Myeloma, the empirical estimate of the survival curve plateaued at around 19% ($\pm 5\%$) for TTR while 30% were estimated to attain the population OS. However, the CrIs for the estimate from the population cure rate models were wide. The analysis showed that a higher disease burden was associated with a higher probability of relapsing for those not cured of Myeloma, based on analysis of TTR. The other covariates: para-protein, beta2 and albumin did not influence the probability of relapsing. In the relative survival model, the probability of attaining the population OS was not associated with log-RD or any of the other observed covariates. The population cure rate models were wrought with fitting problems and required use of weakly informative priors. Even in the ML models, the CIs were very wide implying there was not enough evidence a proportion of the patients would have attained long-term OS. The final part of this chapter compared the effect of the log-RD in the Cox PH, exponential and Weibull models for TTR and OS with those for the TTR and OS in the cure rate models to investigate if there was a change when only those not cured were taken into account. For the TTR, the log-RD effect for those not cured and therefore bound to relapse following treatment in the cure rate models was smaller than in the standard models while there were no obvious differences with respect to the OS.

SEMs were introduced in Chapter 5 with a focus on how latent constructs can be realised from a given number of observed random variables that can be continuous or binary through measurement equations. To ensure model identifiability and also take into account the fact that none of the observed variables were known for certain to be direct measures of the latent construct, only models with the variance of the latent construct fixed at 1 were considered [160]. This was followed by a small simulation study to assess the performance of the extended SEMs with TTE models in the structural model under different censoring schemes. The models performed as expected with no censoring and minimal censoring while with heavy censoring, estimates were not as

close to the true parameters. With increased sample sizes, all estimates were relatively close to the true parameters with good coverage.

Finally, Chapter 6 focused on the application of the proposed SEMs to the Myeloma dataset with a latent variable for tumour growth observed through the three log-transformed continuous random variables, paraprotein, beta2 and albumin. The analysis was applied to a reduced dataset including only those individuals who had available data on each of the markers. The growth rate, on its own and also when fitted with log-RD, was associated with relapse and death following treatment. Estimates in the measurement part for the TTR model were comparable, with the growth rate significantly associated with TTR, OS and the probability of never relapsing after treatment in most cases. The latent growth rate was also associated with each of the 3 observed variables thereby indicating good fit especially in the models with TTR outcomes. When fitted as a covariate with OS as the outcome, estimates relating the observed variables and the growth rate in measurement part were not all significantly associated, and at times the models fitted used both ML and MCMC showed a lack of fit. The use of SEMs to model growth in the cure rate models highlighted its relationship with the probability of being cured and generally, suggested that using the observed variables together in a SEM was better than modelling them individually in the cure rate models of Chapter 4.

7.2 Results in context

Tumour growth rate as an important characteristic in many cancers, has long been associated with adverse outcomes in Myeloma and most of the approaches for modelling the effect of the growth rate on TTE outcomes have involved assuming some distribution of the growth rate [161]. These distributions have been based on serial measurements of the tumour size at times, or through simulations. Other methods of attempting to model the tumour growth rate have involved using surrogate measures of the tumour

burden and growth rate, [162]. In colorectal cancer, SEMs have been used to model the association of a latent variable, oxidative stress, with risk of getting the disease through biomarkers [57]. Other applications of SEMs in cancer have looked at psychosocial adjustment problems [163] in breast cancer, quality of life [164], and other related psychological constructs depicting the mental well-being of cancer patients in general. The presence of biomarkers of tumour growth in some cancers provides an opportunity to use SEMs to model survival outcomes in cancer as illustrated in this thesis. So far, the use of SEMs in cancer has been restricted to measuring non-clinical latent constructs. To effectively use these methods, there would be need to elicit expert knowledge from clinicians on which biomarkers to use in constructing these latent variables depicting important tumour characteristics.

This work has shown that SEMs can be used with TTE outcomes following [56]. However, we exercise caution in the interpretation of the association of the TTE outcome with the growth rate as the selection of the biomarkers used in the measuring this latent construct were not selected based on clinical input, but rather what has been reported in other studies to be the relationship of each of the three biomarkers to the growth rate and RD.

7.3 Limitations

The main limitation of this work was that both the available datasets did not have the requisite plateau from which to estimate both the proportion who will not relapse following treatment or those who will attain the population level OS. The datasets were relatively small with shorter follow-up time compared to the large cohorts that have traditionally been used to estimate the cured proportion [69]. Another challenge was missing data for most of the cytogenetic markers that are known to depict tumour aggressiveness, which made it impossible to use these biomarkers in the SEMs. Finally,

as we did not have expert prior knowledge of how any of the biomarkers relate to tumour growth and therefore TTE outcomes and only non-informative priors were used in the Bayesian models fitted in this analysis, thereby not fully exploiting their power.

7.4 Further work

Several avenues exist for further work. The first would be to work on methods for handling the missing data within the SEMs so that the effect of the latent variable representing tumour growth can be modelled on the TTE outcomes using the already available cytogenetic markers. The sensitivity of the SEMs to the multivariate Normal assumption was only explored by dichotomising paraprotein in the SEM and fitted to the TTR and OS models using ML in Stata. These models can also be implemented in the Bayesian set-up and extended to cure rate models. We mainly used Weibull and Cox PH models for the TTR and OS in the mixture models. More flexible extensions using splines could be implemented for both baseline hazards and PH models [165]. With more input from clinicians, the inclusion of informative priors would help to explore the benefit of a proper Bayesian approach. Due to the limitations seen in the available datasets, these methods could be applied to larger cancer trials that have longer follow-up data.

Appendices

Appendix A

Cure rate models in OpenBUGS

A.1 Mixture model code in OpenBUGS

```
# This code was used to fit the mixture model to estimate the cured proportion
# and model TTR for those not cured
model {
  for (i in 1:n) {
  # Specify the likelihood which includes pi
  L[i]  <- pow(((1-pi[i])*fu[i]), d[i])*pow((pi[i] + (1-pi[i])*Su[i]), (1-d[i]))
  # Usual logistic model for pi
  pi[i] <- exp(alpha0 + alpha1*z[i])/(1 + exp(alpha0 + alpha1*z[i]))
  # Specify model for lambda
  lambda[i] <- exp(beta0 + beta1*x[i])
  # Survivor function for those not cured and bound to relapse
  Su[i] <- exp(-lambda[i]*pow(t[i],alph))
  # Overall survival function that includes pi
  Sov[i] <-  exp((log(pi[i]) - log(pi[i])*Su[i]))
  # The pdf for those not cured
```

```
  fu[i] <- alph*lambda[i]*pow(t[i],(alph-1))*exp(-lambda[i]*pow(t[i],alph))
  # Use Bernoulli trick to maximise likelihood in OpenBUGS
  z[i] <- 1
  z[i] ~ dbern(p[i])
  p[i] <- L[i]/500000

}
  # Priors on the model for pi
alpha0 ~ dnorm(0.0, 0.01)
alpha1 ~ dnorm(0.0, 0.01)
  # Priors on the model for lambda
beta0 ~ dnorm(0.0, 0.01)
beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter
alph ~ dgamma(1,1)
}
```

A.2 Population mixture model code in OpenBUGS

```
# This code was used to fit the population PTC model
model {
  for (i in 1:n) {
  # Specify the log-likelihood which includes the population hazard h[i]
  logLike[i]  <- d[i]*log(h[i] + (1-pi[i])*fu[i]/(pi[i] + (1-pi[i])*Su[i])) + log(pi[i]
                 + (1-pi[i])*Su[i])
  # Usual logistic model for pi
  pi[i] <- exp(alpha0 + alpha1*z[i])/(1 + exp(alpha0 + alpha1*z[i]))
```

```
  # Specify model for lambda
  lambda[i] <- exp(beta0 + beta1*x[i])
  # Survivor function for those not cured and bound to die
  Su[i] <- exp(-lambda[i]*pow(t[i],alph))
  # Overall survival function that includes pi
  Sov[i] <-  exp((log(pi[i]) - log(pi[i])*Su[i]))
  # The pdf for those not cured
  fu[i] <- alph*lambda[i]*pow(t[i],(alph-1))*exp(-lambda[i]*pow(t[i],alph))
  # Use dummy trick to maximise likelihood in OpenBUGS
  dummy[i] <- 0
  dummy[i] ~ dloglik(logLike[i]) # likelihood is exp(logLike[i])
}
  # Priors on the model for pi
alpha0 ~ dnorm(0.0, 0.01)
alpha1 ~ dnorm(0.0, 0.01)
  # Priors on the model for lambda
beta0 ~ dnorm(0.0, 0.01)
beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter
alph ~ dgamma(1,1)
}
```

A.3 Population PTC model code in OpenBUGS

```
# This code was used to fit the population PTC model
model {
  for (i in 1:n) {
```

```
  # Specify the log-likelihood which includes the population hazard h[i]
  logLike[i]  <- d[i]*log(h[i] - log(pi[i])*fu[i]) + (log(pi[i]) - log(pi[i])*Su[i])
  # Usual logistic model for pi
  pi[i] <- exp(alpha0 + alpha1*z[i])/(1 + exp(alpha0 + alpha1*z[i]))
  # Specify model for lambda
  lambda[i] <- exp(beta0 + beta1*x[i])
  # Survivor function for those not cured and bound to die
  Su[i] <- exp(-lambda[i]*pow(t[i],alph))
  # Overall survival function that includes pi
  Sov[i] <-  exp((log(pi[i]) - log(pi[i])*Su[i]))
  # The pdf for those not cured
  fu[i] <- alph*lambda[i]*pow(t[i],(alph-1))*exp(-lambda[i]*pow(t[i],alph))
  # Use dummy trick to maximise likelihood in OpenBUGS
  dummy[i] <- 0
  dummy[i] ~ dloglik(logLike[i]) # likelihood is exp(logLike[i])
}
  # Priors on the model for pi
alpha0 ~ dnorm(0.0, 0.01)
alpha1 ~ dnorm(0.0, 0.01)
  # Priors on the model for lambda
beta0 ~ dnorm(0.0, 0.01)
beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter
alph ~ dgamma(1,1)
}
```

Appendix B

Simulated TTE model with a latent covariate

```
* Code for simulating an Exponential TTE model with a latent variable as a covariate
* in Stata using the package survsim for the TTE model
clear all
set seed 6765327
* Start the program
program simstudy1, rclass
set obs 5000
* Relate the latent contruct to 3 manifest variables assuming it has N(0,1)
gen X = rnormal(0,1)
gen x1 = 1.7 + 0.5*X + rnormal(0, sqrt(2.5))
gen x2 = 3.5 + 0.7*X + rnormal(0, sqrt(0.4))
gen x3 = 3.0 + 1.2*X + rnormal(0, sqrt(2.6))
* Simulate the TTE data
survsim stime, distribution(weibull) lambdas(0.6) gammas(1)covariates(X -0.7)
* Simulate the censoring distribution to esnure approximately 20% are censored
```

```
survsim stimes, distribution(weibull) lambdas(0.15) gammas(1)
* Work out those that are censored
generate died = stime <= stimes
replace stime = stimes if died == 0
* Modify model to fit latent construct on TTE
stset stime, failure(died = 1)
* Call the latent variable G in the TTE model since it was created as X above
gsem (stime <- G, family(weibull, fail(died = 1)))( x1 x2 x3 <- G, variance(G@1))
* Create matrix of estimates
matrix test = r(table)
* Extract specific values from the output to calculate bias and coverage later
* Coefficient of G on time to event
scalar bGTime = test[1,1]
scalar bGTimeSE = test[2,1]
* Constant in the TTE model
scalar TTEcons = test[1,2]
scalar TTEconsSE = test[2,2]
* Coefficient of G on X1
scalar bGx1 = test[1,3]
scalar bx1SE = test[2,3]
* Constant in the model for G on X1
scalar x1cons = test[1,4]
scalar x1consSE = test[2,4]
* Coefficient of G on X2
scalar bGx2 = test[1,5]
scalar bx2SE = test[2,5]
* Constant in the model for G on X2
```

```
scalar x2cons = test[1,6]
scalar x2consSE = test[2,6]
* Coefficient of G on X3
scalar bGx3 = test[1,7]
scalar bx3SE = test[2,7]
* Constant in the model for G on X3
scalar x3cons = test[1,8]
scalar x3consSE = test[2,8]
* The shape parameter estimate
scalar shape = exp(test[1,9])
scalar shapeSE1 = exp(test[5,9])
scalar shapeSE2 = exp(test[6,9])
* Variance of X1
scalar varx1 = test[1,11]
scalar varx1SE = test[2,11]
* Variance of X2
scalar varx2 = test[1,12]
scalar varx2SE = test[2,12]
* Variance of X3
scalar varx3 = test[1,13]
scalar varx3SE = test[2,13]
* End of program
end

* Run the sumilation 1,000 times to create a 1,000 observations for each estimate
simulate bgtime=bGTime bgtse= bGTimeSE TTEcons= TTEcons TTEconsSE= TTEconsSE ///
bGx1= bGx1 bx1SE=bx1SE  x1cons= x1cons x1consSE = x1consSE bGx2 = bGx2 ///
```

```
bx2SE = bx2SE x2cons = x2cons x2consSE = x2consSE bGx3 = bGx3 ///
bx3SE = bx3SE x3cons = x3cons x3consSE = x3consSE shape = shape ///
shapeSE1 = shapeSE1 shapeSE2 = shapeSE2 varx1 = varx1 ///
varx1SE = varx1SE varx2 = varx2 varx2SE = varx2SE ///
varx3 = varx3 varx3SE = varx3SE, reps(1000) nodots nolegend: simstudy1
* Work out the perecentage bias for each parameter in turn. For example for the
* coefficient for the latent variable on the TTE, we have
/* Bias for G on TTE */
generate biasGT = (bgtime - (-0.7))*(100/-0.7)
summarize biasGT, meanonly
display r(mean)
/* Work out the coverage */
generate covGT = (bgtime + invnorm(0.975)*bgtse>-0.7 & bgtime
- invnorm(0.975)*bgtse< -0.7)
tabulate covGT if biasGT!=.
```

Appendix C

OpenBUGS code for fitting Weibull TTE models with SEM extensions

C.1 Bayesian Weibull TTE model with latent tumour growth

```
# This OpenBUGS code is used to fit a Bayesian Weibull TTE model with a latent
# construct as a covariate on the scale parameter
model {
 #Measurement model with logarithms of paraprotein, beta2 and albumin as
  observed
  for (i in 1:n) {
    for (j in 1:3) { y[i,j]~dnorm(mu[i,j], psi[j]) }
    mu[i,1]<-u[1]+tau[1]*omega[i]
    mu[i,2]<-u[2]+tau[2]*omega[i]
    mu[i,3]<-u[3]+tau[3]*omega[i]
    # Specify N(0,1) distribution for the latent growth rate
    omega[i]~dnorm(nu[i], 1)
    # Weibull likelihood function in terms of the pdf and S(t)
```

```
    L[i]  <- pow((alph*lambda[i]*pow(t[i],(alph-1))*exp(-lambda[i]*pow(t[i], alph))),
                 d[i])*pow(exp(-lambda[i]*pow(t[i], alph)),(1 - d[i]))
    # Model for the hazard on the scale parameter
    lambda[i] <- exp(beta0 + beta1*omega[i])
    # A Bernoulli trick to estimate the likelihood function
    z[i] <- 1
    z[i] ~ dbern(p[i])
    p[i] <- L[i]/500000
  } #end of i
  # Prior distributions for the factor loadings
  tau[1]~dnorm(0.0,psi[1])
  tau[2]~dnorm(0.0,psi[2])
  tau[3]~dnorm(0.0,psi[3])
  # Priors for the intercepts and variances
  for (j in 1:3) {
    psi[j]~dgamma(1,1)
    sgm[j]<-1/psi[j]
    u[j]~dnorm(0.0,0.01)
  }
  # Prior for the mean of the latent variable omega
  for (i in 1:n) { nu[i] <- 0}
  # Priors for the model on the scale
  beta0 ~ dnorm(0.0, 0.01)
  beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter in the Weibull model
  alph ~ dgamma(1,1)
} #end of model
```

C.2 Bayesian Cox TTE model with latent tumour growth

```
# This OpenBUGS code is used to fit a Bayesian Cox PH model with a latent
# construct as a covariate on the TTR/OS
model {
    for(i in 1:n) {
    for(j in 1:T) {
    # risk set = 1 if obs.t >= t
    Y[i,j] <- step(obs.t[i] - t[j] + eps)
    # Counting process jump if t[j] <= obs.t < t[j+1]
    dN[i,j] <- Y[i,j]*step(t[j+1] - obs.t[i] - eps)*fail[i]
              }
          }
# Fit the model
for(j in 1:T) {
  for(i in 1:n) {
    dN[i,j]   ~ dpois(Idt[i,j]) # Likelihood
   Idt[i,j] <- Y[i,j]*exp(beta0[j] + beta1*omega[i]) # Intensity
           }
   beta0[j] ~ dnorm(0.0,0.01); # Include the baseline using the Poisson trick
      }
 # Use measurement model with logarithms of paraprotein, beta2 and albumin as
 # observed to get realisations of the latent omega
  for (i in 1:n) {
    for (j in 1:3) { y[i,j]~dnorm(mu[i,j], psi[j]) }
    mu[i,1]<-u[1]+tau[1]*omega[i]
    mu[i,2]<-u[2]+tau[2]*omega[i]
```

```
    mu[i,3]<-u[3]+tau[3]*omega[i]
    # Specify N(0,1) distribution for the latent growth rate
    omega[i]~dnorm(nu[i], 1)
  } #end of i
  # Prior distributions for the factor loadings
  tau[1]~dnorm(0.0,psi[1])
  tau[2]~dnorm(0.0,psi[2])
  tau[3]~dnorm(0.0,psi[3])
  # Priors for the intercepts and variances
  for (j in 1:3) {
    psi[j]~dgamma(1,1)
    sgm[j]<-1/psi[j]
    u[j]~dnorm(0.0,0.01)
  }
  # Prior for the mean of the latent variable omega
  for (i in 1:n) { nu[i] <- 0}
  # Priors for the coefficient in the PH model
  beta1 ~ dnorm(0.0, 0.01)
} #end of model
```

Appendix D

OpenBUGS code for fitting the mixture, PTC and SEM models

D.1 Bayesian Weibull mixture or PTC model with latent tumour growth

```
# This OpenBUGS code is used to fit a Bayesian Weibull TTE model with a latent
# construct from three observed continuous variables on both the cured proportion
# and on the scale parameter for the TTE for those not cured
model {
 #Measurement model with logarithms of paraprotein, beta2 and albumin as
  observed
  for (i in 1:n) {
    for (j in 1:3) { y[i,j]~dnorm(mu[i,j], psi[j]) }
    mu[i,1]<-u[1]+tau[1]*omega[i]
    mu[i,2]<-u[2]+tau[2]*omega[i]
    mu[i,3]<-u[3]+tau[3]*omega[i]
    # Specify N(0,1) distribution for the latent growth rate
```

```
  omega[i]~dnorm(nu[i], 1)
  # Weibull likelihood function with cured proportion of the pdf and S(t) and pi
  L[i]  <- pow(((1-pi[i])*(gamma*lambda[i]*pow(t[i],(gamma-1))*exp(-
           lambda[i]*pow(t[i], gamma)))), d[i])*pow((pi[i]
           + (1-pi[i])*exp(-lambda[i]*pow(t[i], gamma))),(1 - d[i]))
  # Likelihood used when fitting PTC model
  #L[i]  <- pow((-log(pi[i])*alph*lambda[i]*pow(t[i],(alph-1))*exp(-
  #          lambda[i]*pow(t[i], alph))), d[i]) *exp(log(pi[i])-log(pi[i])*exp(-
  #          lambda[i]*pow(t[i], alph)))
  # Logistic model for the cured proportion
  pi[i] <- exp(alpha0 + alpha1*omega[i])/(1 + exp(alpha0 + alpha1*omega[i]))
  # Model for the TTE for those not cured on the scale parameter
  lambda[i] <- exp(beta0 + beta1*omega[i])
  # A Bernoulli trick to estimate the likelihood function
  z[i] <- 1
  z[i] ~ dbern(p[i])
  p[i] <- L[i]/500000
} #end of i
# Prior distributions for the factor loadings
tau[1]~dnorm(0.0,psi[1])
tau[2]~dnorm(0.0,psi[2])
tau[3]~dnorm(0.0,psi[3])
# Priors for the intercepts and variances
for (j in 1:3) {
  psi[j]~dgamma(1,1)
  sgm[j]<-1/psi[j]
  u[j]~dnorm(0.0,0.01)
```

```
  }
  # Prior for the mean of the latent variable omega
  for (i in 1:n) { nu[i] <- 0}
  # Priors on model for pi
  alpha0 ~ dnorm(0.0, 0.01)
  alpha1 ~ dnorm(0.0, 0.01)
  # Priors for the scale model in T
  beta0 ~ dnorm(0.0, 0.01)
  beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter in the Weibull model
  gamma ~ dgamma(1,1)
 } #end of model
```

D.2 Bayesian Weibull population mixture or PTC models with latent tumour growth

```
# This OpenBUGS code is used to fit a Bayesian Weibull TTE model with a latent
# construct from three observed continuous variables on both the cured proportion
# and on the scale parameter for the TTE for those bound to down
model {
 #Measurement model with logarithms of paraprotein, beta2 and albumin as
  observed
  for (i in 1:n) {
    for (j in 1:3) { y[i,j]~dnorm(mu[i,j], psi[j]) }
    mu[i,1]<-u[1]+tau[1]*omega[i]
    mu[i,2]<-u[2]+tau[2]*omega[i]
    mu[i,3]<-u[3]+tau[3]*omega[i]
```

```
  # Specify N(0,1) distribution for the latent growth rate
  omega[i]~dnorm(nu[i], 1)
  # Weibull likelihood function with cured proportion of the pdf and S(t) and pi
  # with population hazard rate h[i] for each individual
  L[i]  <- pow((h[i] + (1-pi[i])*(alph*lambda[i]*pow(t[i],(alph-1))*exp(-
         lambda[i]*pow(t[i], alph)))/(pi[i] + (1-pi[i])*exp(-lambda[i]*pow(t[i],
     alph)))), d[i]) *pow((pi[i] + (1-pi[i])*exp(-lambda[i]*pow(t[i], alph))),1-d[i])

  # Likelihood used when fitting PTC model
  #L[i]  <- pow((h[i] - log(pi[i])*alph*lambda[i]*pow(t[i],(alph-1))*exp(-
  #          lambda[i]*pow(t[i],alph))), d[i])*exp((log(pi[i]) - log(pi[i])*exp(-
  #           lambda[i]*pow(t[i],alph))))
  # Logistic model for the cured proportion
  pi[i] <- exp(alpha0 + alpha1*omega[i])/(1 + exp(alpha0 + alpha1*omega[i]))
  # Model for the TTE for those not cured on the scale parameter
  lambda[i] <- exp(beta0 + beta1*omega[i])
  # A Bernoulli trick to estimate the likelihood function
  z[i] <- 1
  z[i] ~ dbern(p[i])
  p[i] <- L[i]/500000
} #end of i
# Prior distributions for the factor loadings
tau[1]~dnorm(0.0,psi[1])
tau[2]~dnorm(0.0,psi[2])
tau[3]~dnorm(0.0,psi[3])
# Priors for the intercepts and variances
for (j in 1:3) {
```

```
    psi[j]~dgamma(1,1)
    sgm[j]<-1/psi[j]
    u[j]~dnorm(0.0,0.01)
  }
  # Prior for the mean of the latent variable omega
  for (i in 1:n) { nu[i] <- 0}
  # Priors on model for pi
  alpha0 ~ dnorm(0.0, 0.01)
  alpha1 ~ dnorm(0.0, 0.01)
  # Priors for the scale model in T
  beta0 ~ dnorm(0.0, 0.01)
  beta1 ~ dnorm(0.0, 0.01)
  # Prior for the shape parameter in the Weibull model
  alph ~ dgamma(1,1)
} #end of model
```

Appendix E

Assessment of model fits in Myeloma using scaled Schoenfeld residuals

E.1 Assessing the multivariable Cox PH model for TTR in Myeloma

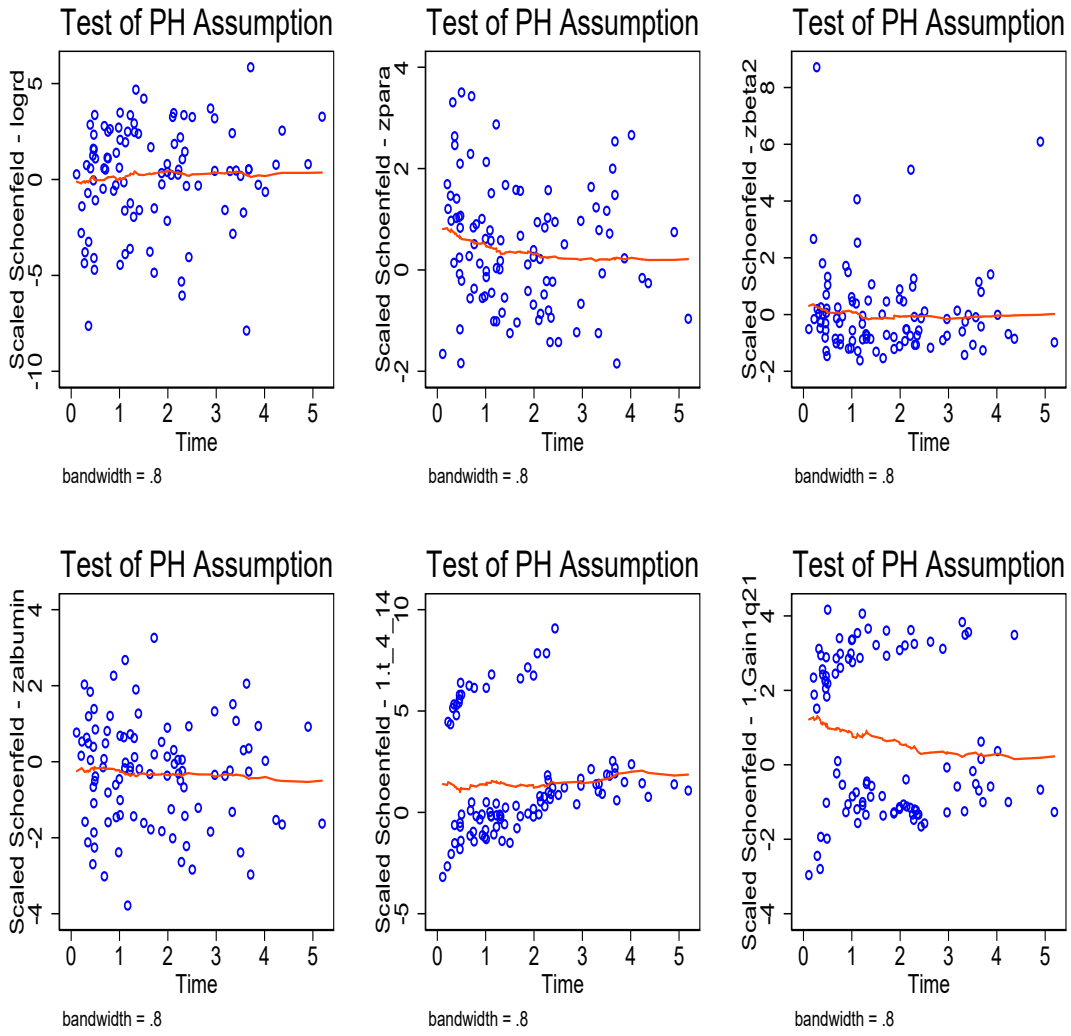


Figure E.1: Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for TTR with 125 patients

E.2 Assessing the multivariable Cox PH model for OS in Myeloma

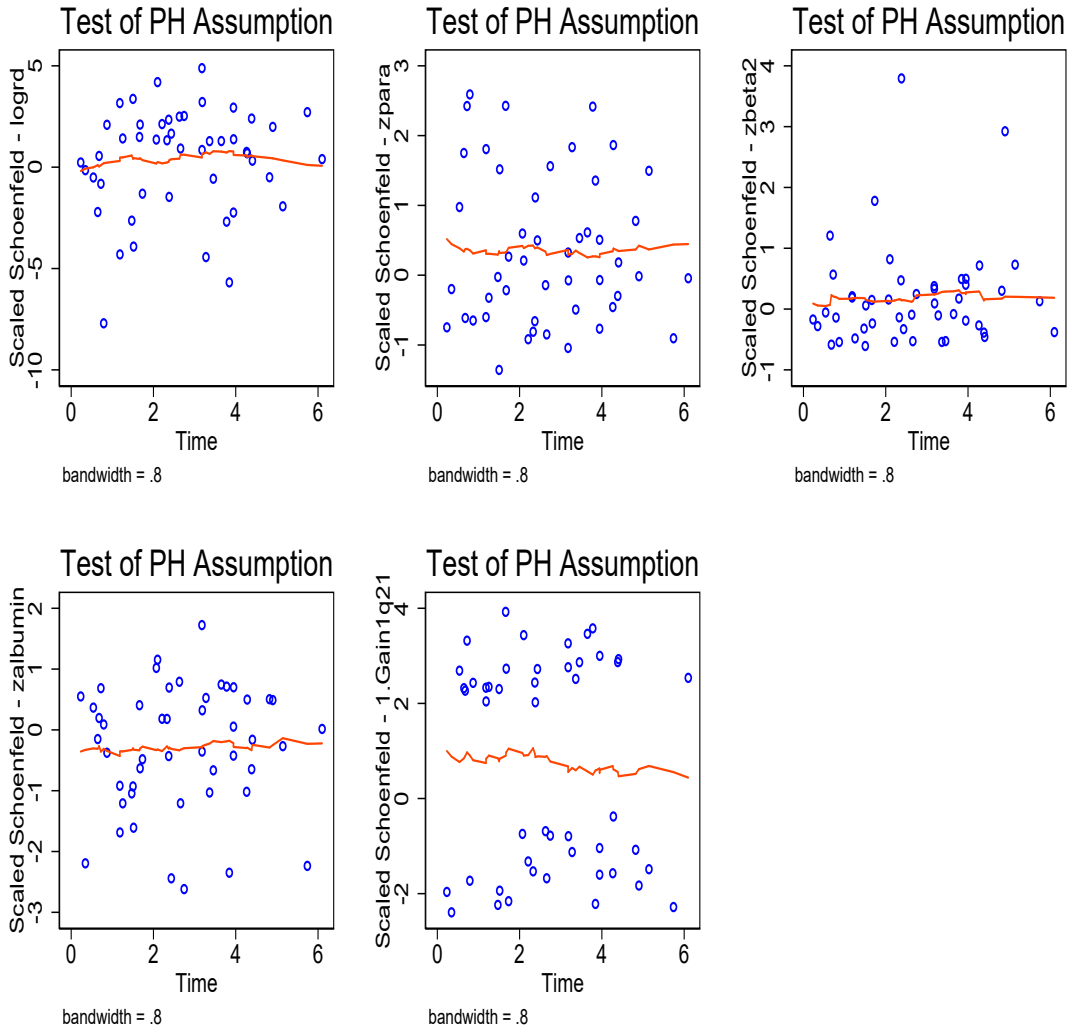


Figure E.2: Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for OS with 125 patients

Appendix F

Assessment of model fits in CLL using scaled Schoenfeld residuals

F.1 Assessing the multivariable Cox PH model for TTR in CLL

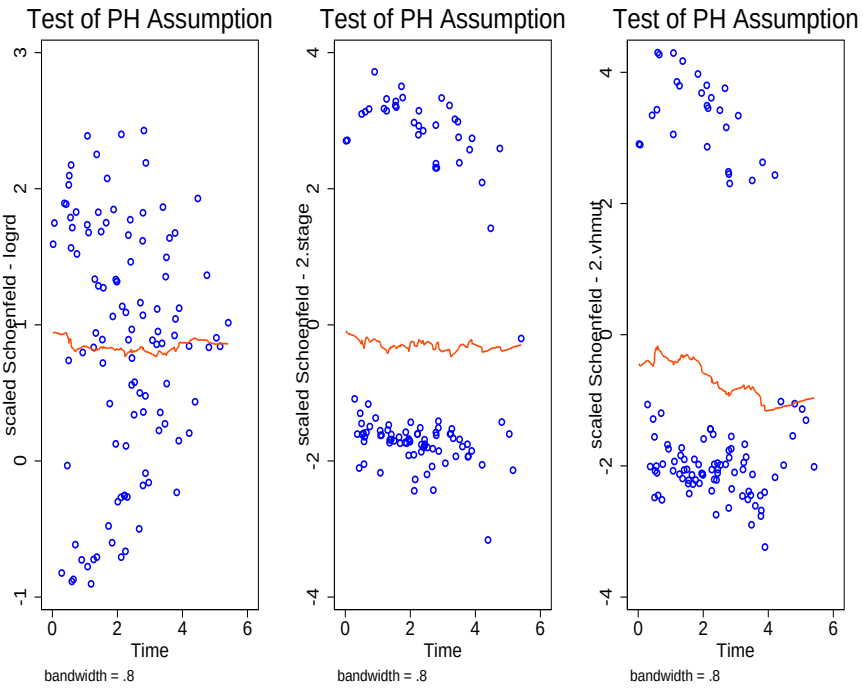


Figure F.1: Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for TTR in CLL fitted to data from 286 patients

F.2 Assessing the multivariable Cox PH model for OS in CLL

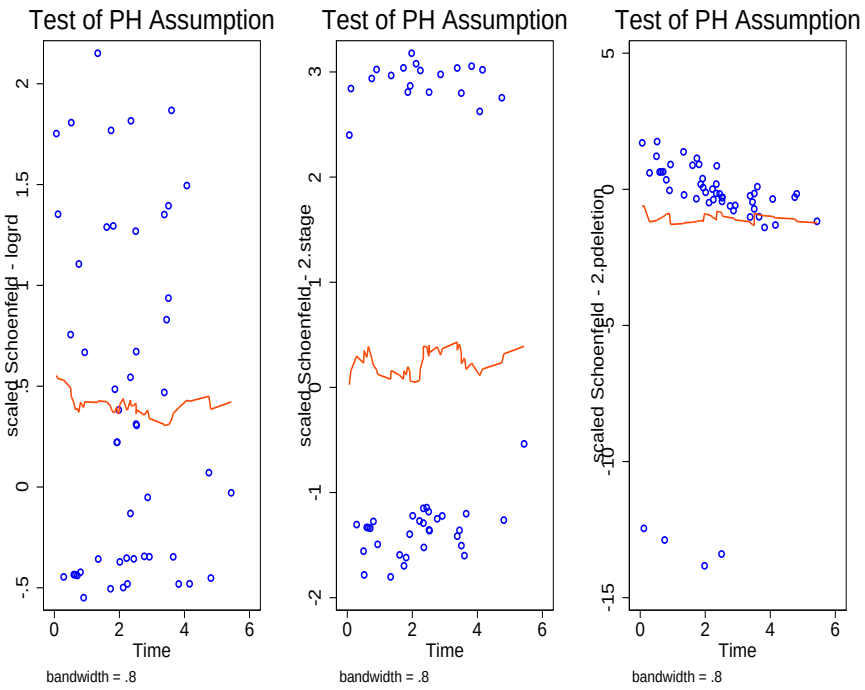


Figure F.2: Plots of the Schoenfeld residuals for each of the covariates in the final multivariable Cox PH model for OS with 301 patients

Bibliography

[1] Cancer statistics for the UK;. Accessed: 2019-03-20. https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk.

[2] About blood cancer;. Accessed: 2019-03-20. https://www.dkms.org.uk/en/about-blood-cancer.

[3] Comen E, Morris PG, Norton L. Translating mathematical modeling of tumor growth patterns into novel therapeutic approaches for breast cancer. Journal of mammary gland biology and neoplasia. 2012;17(3-4):241–249.

[4] Ghossein RA, Rosai J. Polymerase chain reaction in the detection of micrometastases and circulating tumor cells. Cancer: Interdisciplinary International Journal of the American Cancer Society. 1996;78(1):10–16.

[5] Ghossein RA, Bhattacharya S, Rosai J. Molecular detection of micrometastases and circulating tumor cells in solid tumors. Clinical Cancer Research. 1999;5(8):1950–1960.

[6] van Dongen JJ, Seriu T, Panzer-Grümayer ER, Biondi A, Pongers-Willemse MJ, Corral L, et al. Prognostic value of minimal residual disease in acute lymphoblastic leukaemia in childhood. The Lancet. 1998;352(9142):1731 – 1738.

[7] Böttcher S, Ritgen M, Fischer K, Stilgenbauer S, Busch RM, Fingerle-Rowson G, et al. Minimal residual disease quantification is an independent predictor of progression-free and overall survival in chronic lymphocytic leukemia: a multivariate analysis from the randomized GCLLSG CLL8 trial. Journal of clinical oncology. 2012;30(9):980–988.

[8] Heymach J, Krilov L, Alberg A, Baxter N, Chang SM, Corcoran RB, et al. Clinical cancer advances 2018: annual report on progress against cancer from the American Society of Clinical Oncology. Journal of Clinical Oncology. 2018;36(10):1020–1044.

[9] Woods LM, Rachet B, Lambert PC, Coleman MP. 'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis. Annals of Oncology. 2009;20(8):1331–1336.

[10] Sargent D, Sobrero A, Grothey A, O'Connell MJ, Buyse M, Andre T, et al. Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. Journal of Clinical Oncology. 2009;27(6):872.

[11] Verdecchia A, De Angelis R, Capocaccia R, Sant M, Micheli A, Gatta G, et al. The cure for colon cancer: results from the EUROCARE study. International Journal of Cancer. 1998;77(3):322–329.

[12] Andrae B, Andersson TML, Lambert PC, Kemetli L, Silfverdal L, Strander B, et al. Screening and cervical cancer cure: population based cohort study. BMJ. 2012;344. Available from: https://www.bmj.com/content/344/bmj.e900.

[13] Dubecz A, Gall I, Solymosi N, Schweigert M, Peters JH, Feith M, et al. Temporal Trends in Long-Term Survival and Cure Rates in Esophageal Cancer: A SEER Database Analysis. Journal of Thoracic Oncology. 2012;7(2):443 –

447. Available from: http://www.sciencedirect.com/science/article/pii/S1556086415332512.

[14] Francisci S, Capocaccia R, Grande E, Santaquilani M, Simonetti A, Allemani C, et al. The cure of cancer: a European perspective. European Journal of Cancer. 2009;45(6):1067–1079.

[15] Ravi P, Kumar SK, Cerhan JR, Maurer MJ, Dingli D, Ansell SM, et al. Defining cure in multiple myeloma: a comparative study of outcomes of young individuals with myeloma and curable hematologic malignancies. Blood cancer journal. 2018;8(3):26.

[16] Dickman PW, Adami HO. Interpreting trends in cancer patient survival. Journal of internal medicine. 2006;260(2):103–117.

[17] Adelian R, Jamali J, Zare N, Ayatollahi S, Pooladfar G, Roustaei N. Comparison of Cox's regression model and parametric models in evaluating the prognostic factors for survival after liver transplantation in Shiraz during 2000–2012. International journal of organ transplantation medicine. 2015;6(3):119.

[18] Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society Series B (Methodological). 1949;11(1):15–53.

[19] Johnson P, Greiner W, Al-Dakkak I, Wagner S. Which metrics are appropriate to describe the value of new cancer therapies? BioMed research international. 2015;2015.

[20] Tai P, Yu E, Cserni G, Vlastos G, Royce M, Kunkler I, et al. Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database. BMC Cancer. 2005 May;5(1):48. Available from: https://doi.org/10.1186/1471-2407-5-48.

[21] Barlogie B, Mitchell A, van Rhee F, Epstein J, Morgan GJ, Crowley J. Curing myeloma at last: defining criteria and providing the evidence. Blood. 2014;124(20):3043–3051. Available from: http://www.bloodjournal.org/content/124/20/3043.

[22] Othus M, Barlogie B, LeBlanc ML, Crowley JJ. Cure Models as a Useful Statistical Tool for Analyzing Survival. Clinical Cancer Research. 2012;.

[23] Katodritou E, Papadaki S, Konstantinidou P, Terpos E. Is it possible to cure myeloma without allogeneic transplantation? Transfusion and Apheresis Science. 2016;54(1):63 – 70.

[24] Wang D. Maller RA, Zhou X, Survival analysis with long-term survivors. STATISTICAL METHODS IN MEDICAL RESEARCH. 2000;9:520–520.

[25] Cai C, Zou Y, Peng Y, Zhang J. smcure: An R-Package for estimating semiparametric mixture cure models. Computer methods and programs in biomedicine. 2012;108(3):1255–1260.

[26] Sy JP, Taylor JM. Estimation in a Cox proportional hazards cure model. Biometrics. 2000;56(1):227–236.

[27] Sposto R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. Statistics in medicine. 2002;21(2):293–312.

[28] Yu B, Tiwari RC, Cronin KA, Feuer EJ. Cure fraction estimation from the mixture cure models for grouped survival data. Statistics in Medicine. 2004;23(11):1733–1747.

[29] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics. 2007;8(3):576–594.

[30] De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in Medicine. 1999;18(4):441–454.

[31] Andersson TM, Dickman PW, Eloranta S, Lambert PC. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. BMC Medical Research Methodology. 2011 Jun;11(1):96.

[32] Partridge SC, Gibbs JE, Lu Y, Esserman LJ, Tripathy D, Wolverton DS, et al. MRI measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival. American Journal of Roentgenology. 2005;184(6):1774–1781.

[33] Winer-Muram HT, Jennings SG, Tarver RD, Aisen AM, Tann M, Conces DJ, et al. Volumetric growth rate of stage I lung cancer prior to treatment: serial CT scanning. Radiology. 2002;223(3):798–805.

[34] Steel G, Lamerton L. The growth rate of human tumours. British journal of cancer. 1966;20(1):74.

[35] Collins VP. Observation on growth rates of human tumors. Am J Roentgenol. 1956;76:988–1000.

[36] Spratt Jr JS, Spratt TL. Rates of growth of pulmonary metastases and host survival. Annals of Surgery. 1964;159(2):161.

[37] Spratt JS. The lognormal frequency distribution and human cancer. Journal of Surgical Research. 1969;9(3):151–157.

[38] Laird AK. Dynamics of tumour growth. British journal of cancer. 1964;18(3):490.

[39] Simon R, Norton L. The Norton-Simon hypothesis: designing more effective and less toxic chemotherapeutic regimens. Nature Reviews Clinical Oncology. 2006;3(8):406.

[40] Gregory WM, Richards MA, Slevin ML, Souhami RL. A mathematical model relating response durations to amount of subclinical resistant disease. Cancer research. 1991;51(4):1210–1216.

[41] Gregory WM, Twelves CJ, Bell R, Smye SW, Howard DR, Coleman RE, et al. Characterizing and quantifying the effects of breast cancer therapy using mathematical modeling. Breast cancer research and treatment. 2016;155(2):303–311.

[42] Paulino CDM, de Bragança Pereira CA. On identifiability of parametric statistical models. Journal of the Italian Statistical Society. 1994;3(1):125–151.

[43] Day RS, Shackney SE, Peters WP. The analysis of relapse-free survival curves: implications for evaluating intensive systemic adjuvant treatment regimens for breast cancer. British journal of cancer. 2005;92(1):47.

[44] Norton L. A Gompertzian model of human breast cancer growth. Cancer research. 1988;48(24 Part 1):7067–7071.

[45] Vargo-Gogola T, Rosen JM. Modelling breast cancer: one size does not fit all. Nature Reviews Cancer. 2007;7(9):659.

[46] Byrne H, Alarcon T, Owen M, Webb S, Maini P. Modelling aspects of cancer dynamics: a review. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. 2006;364(1843):1563–1578.

[47] Zumkeller W. IGFs and IGFBPs: surrogate markers for diagnosis and surveillance of tumour growth? Molecular Pathology. 2001;54(5):285.

[48] Lamerz R. Role of tumour markers, cytogenetics. Annals of oncology. 1999;10(suppl_4):S145–S149.

[49] Kumar S, Paiva B, Anderson KC, Durie B, Landgren O, Moreau P, et al. International Myeloma Working Group consensus criteria for response and minimal residual disease assessment in multiple myeloma. The Lancet Oncology. 2016;17(8):e328–e346.

[50] Rawstron AC, Gregory WM, de Tute RM, Davies FE, Bell S, Drayson MT, et al. Minimal residual disease in myeloma by flow cytometry: independent prediction of survival benefit per log reduction. Blood. 2015;p. blood–2014.

[51] Munshi NC, Avet-Loiseau H, Rawstron AC, Owen RG, Child JA, Thakurta A, et al. Association of minimal residual disease with superior survival outcomes in patients with multiple myeloma: a meta-analysis. JAMA oncology. 2017;3(1):28–35.

[52] Anderson JC, Gerbing DW. Structural equation modeling in practice: A review and recommended two-step approach. Psychological bulletin. 1988;103(3):411.

[53] Clogg CC. Structural Equations with Latent Variables. JSTOR; 1991.

[54] Finney SJ, DiStefano C. Non-normal and categorical data in structural equation modeling. Structural equation modeling: A second course. 2006;10(6):269–314.

[55] Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika. 1969;34(2):183–202.

[56] Larsen K. The Cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. Biometrics. 2005;61(4):1049–1055.

[57] Eldridge RC, Goodman M, Bostick RM, Fedirko V, Gross M, Thyagarajan B, et al. A Novel Application of Structural Equation Modeling Estimates the As-

sociation between Oxidative Stress and Colorectal Adenoma. Cancer Prevention Research. 2018;11(1):52–58.

[58] Cohen P, Cohen J, Teresi J, Marchi M, Velez CN. Problems in the measurement of latent variables in structural equations causal models. Applied Psychological Measurement. 1990;14(2):183–196.

[59] Hollander M, Wolfe DA. Nonparametric statistical methods. 1999;.

[60] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187–220.

[61] Elandt R, Elandt-Johnson RC, Johnson NL. Survival models and data analysis. vol. 110. John Wiley & Sons; 1980.

[62] Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics. 1982;p. 1041–1046.

[63] National life tables: UK;. Accessed: 2019-03-20. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables.

[64] Cowles MK. Modelling Survival Data in Medical Research. Journal of the American Statistical Association. 2004;99(467):905–907.

[65] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research. 2010;1(4):274.

[66] David CR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society. 1972;34:187–220.

[67] Cox DR. Partial likelihood. Biometrika. 1975;62(2):269–276.

[68] Cox D, Oakes D. Analysis of survival data. Chapman and Hall, London; 1984.

[69] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. Biostatistics. 2007;8(3):576–594.

[70] Bilmes JA, et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. 1998;4(510):126.

[71] Stata;. https://www.stata.com/.

[72] The R Project for Statistical Computing;. Accessed: 2019-06-12. https://www.r-project.org/.

[73] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Selected papers of hirotugu akaike. Springer; 1998. p. 199–213.

[74] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977;p. 1–38.

[75] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977;39(1):1–22. Available from: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

[76] Van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov Chain Monte–Carlo sampling. Psychonomic bulletin & review. 2018;25(1):143–154.

[77] Bernardo JM, Smith AF. Bayesian theory. vol. 405. John Wiley & Sons; 2009.

[78] Ntzoufras I. Bayesian modeling using WinBUGS. vol. 698. John Wiley & Sons; 2011.

[79] Link WA, Eaton MJ. On thinning of chains in MCMC. Methods in ecology and evolution. 2012;3(1):112–115.

[80] Kalbfleisch JD. Non-parametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society: Series B (Methodological). 1978;40(2):214–221.

[81] Leukemia Bayesian Cox regression in BUGS;. http://www.openbugs.net/Examples/Leuk.html.

[82] Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239–241.

[83] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–526.

[84] Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika. 1990;77(1):147–160.

[85] Cox DR, Snell EJ. A general definition of residuals. Journal of the Royal Statistical Society Series B (Methodological). 1968;p. 248–275.

[86] Bradburn M, Clark T, Love S, Altman D. Survival analysis Part III: multivariate data analysis–choosing a model and assessing its adequacy and fit. British journal of cancer. 2003;89(4):605.

[87] Wileyto EP, Li Y, Chen J, Heitjan DF. Assessing the fit of parametric cure models. Biostatistics. 2013;14(2):340–350.

[88] Hsu WW, Todem D, Kim K. A sup-score test for the cure fraction in mixture models for long-term survivors. Biometrics. 2016;72(4):1348–1357.

[89] Müller UU, Van Keilegom I. Goodness-of-fit tests for the cure rate in a mixture cure model. Biometrika. 2019;106(1):211–227.

[90] Lambert PC. Modeling of the cure fraction in survival studies. The Stata Journal. 2007;7(3):351–375.

[91] Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. Statistics in medicine. 2009;28(25):3049–3067.

[92] Plummer M, et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. vol. 124. Vienna, Austria.; 2003. p. 10.

[93] Gelman A, Lee D, Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization. Journal of Educational and Behavioral Statistics. 2015;40(5):530–543.

[94] Gelman A, Rubin DB, et al. Inference from iterative simulation using multiple sequences. Statistical science. 1992;7(4):457–472.

[95] Gelman A, Rubin DB. A single series from the Gibbs sampler provides a false sense of security. Bayesian statistics. 1992;4:625–631.

[96] Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica. 1996;p. 733–760.

[97] Gourieroux C, Holly A, Monfort A. Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. Econometrica: journal of the Econometric Society. 1982;p. 63–80.

[98] Berg A, Meyer R, Yu J. Deviance information criterion for comparing stochastic volatility models. Journal of Business & Economic Statistics. 2004;22(1):107–120.

[99] Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987;52(3):345–370.

[100] Neath AA, Cavanaugh JE. The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(2):199–203.

[101] Wilberg MJ, Bence JR. Performance of deviance information criterion model selection in statistical catch-at-age analysis. Fisheries Research. 2008;93(1-2):212–221.

[102] Spiegelhalter DJ, Best NG, Carlin BP, Linde A. The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014;76(3):485–493.

[103] Morgan GJ, Davies FE, Gregory WM, Cocks K, Bell SE, Szubert AJ, et al. First-line treatment with zoledronic acid as compared with clodronic acid in multiple myeloma (MRC Myeloma IX): a randomised controlled trial. The Lancet. 2010;376(9757):1989–1999.

[104] Rajan A, Rajkumar SV. Interpretation of cytogenetic results in multiple myeloma for clinical practice. Blood cancer journal. 2015;5(10):e365.

[105] Camilli G, Hopkins KD. Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. Psychological Bulletin. 1978;85(1):163.

[106] Fitzmaurice G. Confounding: regression adjustment. Nutrition. 2006;22(5):581.

[107] Howard D, Munir T, McParland L, Rawstron A, Milligan D, Schuh A, et al. Results of the randomized phase IIB ARCTIC trial of low-dose rituximab in previously untreated CLL. Leukemia. 2017;31(11):2416.

[108] Munir T, Howard D, McParland L, Pocock C, Rawstron A, Hockaday A, et al. Results of the randomized phase IIB ADMIRE trial of FCR with or without mitoxantrone in previously untreated CLL. Leukemia. 2017;31(10):2085.

[109] Stages of chronic lymphocytic leukemia;. Accessed: 2019-09-07. https://www.cancer.ca/en/cancer-information/cancer-type/leukemia-chronic-lymphocytic-cll/staging/?region=on.

[110] Yu L, Kim HT, Kasar S, Benien P, Du W, Hoang K, et al.. Comprehensive genetic characterization of 17p deleted CLL identifies predictors of overall survival. Am Soc Hematology; 2015.

[111] Little RJ, Rubin DB. The analysis of social science data with missing values. Sociological Methods & Research. 1989;18(2-3):292–326.

[112] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine. 2010;29(28):2920–2931.

[113] White IR, Royston P. Imputing missing covariate values for the Cox model. In: Statistics in medicine; 2009. .

[114] Rubin DB. Multiple imputation for nonresponse in surveys. vol. 81. John Wiley & Sons; 2004.

[115] White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377–399.

[116] Grace-Martin K. Multiple imputation in a nutshell. Retrieved from. 2016;.

[117] Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging themes in epidemiology. 2017;14(1):8.

[118] Cattle BA, Baxter PD, Greenwood DC, Gale CP, West RM. Multiple imputation for completion of a national clinical audit dataset. Statistics in medicine. 2011;30(22):2736–2753.

[119] White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Computational statistics & data analysis. 2010;54(10):2267–2275.

[120] Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. Statistics in medicine. 2018;37(25):3661–3678.

[121] Schafer JL. Multiple imputation: a primer. Statistical methods in medical research. 1999;8(1):3–15.

[122] Shah V, Sherborne AL, Walker BA, Johnson DC, Boyle EM, Ellis S, et al. Prediction of outcome in newly diagnosed myeloma: a meta-analysis of the molecular profiles of 1905 trial patients. Leukemia. 2018;32(1):102–110.

[123] Rozovski U, Keating MJ, Estrov Z. Why is the immunoglobulin heavy chain gene mutation status a prognostic indicator in chronic lymphocytic leukemia? Acta haematologica. 2018;140(1):51–54.

[124] Yu L, Kim HT, Kasar SN, Benien P, Du W, Hoang K, et al. Survival of Del17p CLL depends on genomic complexity and somatic mutation. Clinical Cancer Research. 2017;23(3):735–745.

[125] Maller RA, Zhou X. Survival Analysis with Long-Term Survivors. Wiley Series in Child Care and Protection. Wiley; 1996. Available from: `https://books.google.co.uk/books?id=V3hFAAAAYAAJ`.

[126] Yin G, Ibrahim JG. Cure rate models: a unified approach. Canadian Journal of Statistics. 2005;33(4):559–570.

[127] Kutal D, Qian L. A Non-Mixture Cure Model for Right-Censored Data with Fréchet Distribution. Stats. 2018;1(1):176–188.

[128] Amico M, Van Keilegom I. Cure models in survival analysis. Annual Review of Statistics and Its Application. 2018;5:311–342.

[129] Kutal D, Qian L. A Non-Mixture Cure Model for Right-Censored Data with Fréchet Distribution. Stats. 2018 11;1:176–188.

[130] De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. Statistics in medicine. 1999;18(4):441–454.

[131] Lunn D, Jackson C, Best N, Spiegelhalter D, Thomas A. The BUGS book: A practical introduction to Bayesian analysis. Chapman and Hall/CRC; 2012.

[132] Li CS, Taylor JM, Sy JP. Identifiability of cure models. Statistics & Probability Letters. 2001;54(4):389–395.

[133] Kim S, Choi S, Yoon JH, Kim Y, Lee S, Park T. Drug response prediction model using a hierarchical structural component modeling method. BMC bioinformatics. 2018;19(9):33–44.

[134] Baltar VT, Xun WW, Johansson M, Ferrari P, Chuang SC, Relton C, et al. A structural equation modelling approach to explore the role of B vitamins and immune markers in lung cancer risk. European journal of epidemiology. 2013;28(8):677–688.

[135] Asparouhov T, Masyn K, Muthen B. Continuous time survival in latent variable models. In: Proceedings of the Joint Statistical Meeting in Seattle; 2006. p. 180–187.

[136] McCurdy SR, Molinaro A, Pachter L. A latent variable model for survival time prediction with censoring and diverse covariates. arXiv preprint arXiv:170606995. 2017;.

[137] Muthén B, Masyn K. Discrete-time survival mixture analysis. Journal of Educational and Behavioral statistics. 2005;30(1):27–58.

[138] Song XY, Lee SY. A tutorial on the Bayesian approach for analyzing structural equation models. Journal of Mathematical Psychology. 2012;56(3):135–148.

[139] Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine. 2005;24(11):1713–1723.

[140] Anderson TW. An introduction to multivariate statistical analysis. Wiley New York; 1962.

[141] Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Chapman and Hall/CRC; 2004.

[142] Lee SY, Song XY. Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences. John Wiley & Sons; 2012.

[143] Robert CP, Elvira V, Tawn N, Wu C. Accelerating MCMC algorithms. WIREs Computational Statistics. 2018;10(5):e1435. Available from: `https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1435`.

[144] Hamilton LC. Stata Structural Equation Modeling Reference ManualVRelease 12. College Station, TX: StataCorp LP. 2011;.

[145] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika. 1974;61(2):215–231.

[146] Gonzalez R, Griffin D. Testing parameters in structural equation modeling: Every" one" matters. Psychological Methods. 2001;6(3):258.

[147] Suhr D. The basics of structural equation modeling. Presented: Irvine, CA, SAS User Group of the Western Region of the United States (WUSS). 2006;.

[148] Crowther MJ, Lambert PC. Simulating complex survival data. The Stata Journal. 2012;12(4):674–687.

[149] Crowther MJ. SURVSIM: Stata module to simulate complex survival data; 2011. Statistical Software Components, Boston College Department of Economics. Available from: https://ideas.repec.org/c/boc/bocode/s457317.html.

[150] Huber C, et al. Generalized structural equation modeling in Stata. In: Italian Stata Users' Group Meetings 2013. 06. Stata Users Group; 2013. .

[151] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25(24):4279–4292.

[152] Wolf EJ, Harrington KM, Clark SL, Miller MW. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. Educational and psychological measurement. 2013;73(6):913–934.

[153] Brooks SP, Gelman A. General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics. 1998;7(4):434–455. Available from: https://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474787.

[154] Fraser GE, Stram DO. Regression calibration when foods (measured with error) are the variables of interest: markedly non-Gaussian data with many zeroes. American journal of epidemiology. 2012;175(4):325–331.

[155] Escalante A, Del Rincón I, Cornell JE. Latent variable approach to the measurement of physical disability in rheumatoid arthritis. Arthritis Care & Research. 2004;51(3):399–407.

[156] Galimberti S, Benedetti E, Morabito F, Papineschi F, Callea V, Fazzi R, et al. Prognostic role of minimal residual disease in multiple myeloma patients after non-

myeloablative allogeneic transplantation. Leukemia research. 2005;29(8):961–966.

[157] Sawyer JR. The prognostic significance of cytogenetics and molecular profiling in multiple myeloma. Cancer genetics. 2011;204(1):3–12.

[158] Kröber A, Seiler T, Benner A, Bullinger L, Brückle E, Lichter P, et al. VH mutation status, CD38 expression level, genomic aberrations, and survival in chronic lymphocytic leukemia. Blood. 2002;100(4):1410–1416.

[159] Badoux XC, Keating M, O'Brien S, Ferrajoli A, Burger JA, Faderl S, et al.. Patients with Relapsed CLL and 17p Deletion by FISH Have Very Poor Survival Outcomes. Am Soc Hematology; 2009.

[160] Bentler PM, Chou CP. Practical issues in structural modeling. Sociological Methods & Research. 1987;16(1):78–117.

[161] Sullivan PW, Salmon SE. Kinetics of tumor growth and regression in IgG multiple myeloma. The Journal of clinical investigation. 1972;51(7):1697–1708.

[162] Mai EK, Hielscher T, Kloth JK, Merz M, Shah S, Hillengass M, et al. Association between magnetic resonance imaging patterns and baseline disease features in multiple myeloma: analyzing surrogates of tumour mass and biology. European radiology. 2016;26(11):3939–3948.

[163] Vos PJ, Garssen B, Visser AP, Duivenvoorden HJ, de Haes HC. Early stage breast cancer: explaining level of psychosocial adjustment using structural equation modeling. Journal of Behavioral Medicine. 2004;27(6):557–580.

[164] Nuamah IF, Cooley ME, Fawcett J, McCorkle R. Testing a theory for health-related quality of life in cancer patients: A structural equation approach. Research in nursing & health. 1999;22(3):231–242.

[165] Royston P, Lambert PC, et al. Flexible parametric survival analysis using Stata: beyond the Cox model. 2011;.