Comparative Analysis of Joint Modeling and Time-Dependent Cox Models: Insights from a Longitudinal Study on HIV Patients' Survival and CD4 Cell Count Dynamics

Ryan Wei

Abstract

This report examines the efficacy of Joint Modeling (JM) versus time-dependent Cox models in analyzing survival and CD4 cell count data from 467 HIV-infected patients. By integrating longitudinal CD4 cell counts with survival outcomes, the JM approach offers a nuanced analysis, overcoming challenges associated with traditional Cox models. The comparison reveals JM's superiority in providing detailed insights into the relationship between CD4 trajectories and patient survival, emphasizing its value in personalized healthcare and advancing the field of survival analysis in medical research.

Keywords: longitudinal data analysis, survival data, time-dependent variables, joint modelling, HIV, CD4

1 Introduction

Joint modeling has emerged as a compelling and increasingly popular approach in statistical and medical research, as highlighted in key works by Rizopoulos and Lesaffre (2014), Tsiatis and Davidian (2004). These models are particularly valuable in longitudinal studies aimed at linking a time-dependent response with an event time outcome. Generally, joint models are pivotal in two primary scenarios.

Firstly, they are crucial when assessing the relationship between the risk of an event and a time-varying covariate, especially when the covariate in question is endogenous, as discussed by Kalbfleisch and Prentice (2002). Traditional methods like the time-dependent Cox model (Therneau and Grambsch (2000)) may not be the most effective in these instances. Endogenous covariates, such as biomarkers or other patient-specific measurements taken during a study, have a direct link to the event's occurrence. Modeling the joint distribution of these covariates and the event processes allows for a more nuanced understanding of their interrelation, leading to more precise estimations of their association.

Secondly, joint models are indispensable in dealing with incomplete data scenarios. This is particularly relevant when the likelihood of missing data is dependent on unobserved longitudinal responses. As noted by Little and Rubin (2019) and Molenberghs and Kenward (2007), valid inferences in such situations require a model that jointly considers the distributions of the longitudinal data and the missingness mechanisms. The R package JMbayes, for example, stands as a testament to the advancements in this field, offering sophisticated tools for such analyses.

Joint models typically consist of two main components: a longitudinal submodel and a survival submodel (Ibrahim et al. (2010)). The longitudinal submodel, often a linear mixed-effects model, is used to describe the trajectory of the repeated measurements over time. The survival submodel, usually a proportional hazards model, is used to analyze the time-to-event data. These two models are linked by shared random effects and/or by including the predicted values of the longitudinal outcome in the survival submodel.

This methodology allows for a more comprehensive and nuanced understanding of the data, accounting for the complexities inherent in longitudinal studies. It provides insights into how changes over time in certain variables may influence the risk of an event, which is invaluable

in fields like medicine, where it can inform treatment decisions and improve patient outcomes.

In this report, we will employ joint modeling techniques to analyze a dataset derived from a study conducted by Goldman et al. (1996), which is a randomized clinical trial comparing the efficacy and safety of Didanosine and Zalcitabine in HIV patients. This trial, crucial in the landscape of antiretroviral therapy, specifically targeted patients who had either failed or exhibited intolerance towards zidovudine (AZT) treatment. This methodological choice is motivated by the inherent structure of the data, where the longitudinal measurements (such as CD4 counts and other biomarkers) are expected to be closely intertwined with the time-to-event outcomes (like survival or drug efficacy). By applying joint models, we aim to unravel the complex interplay between these two aspects, providing deeper insights into the effectiveness and safety profiles of Didanosine and Zalcitabine in the context of HIV treatment.

2 Method

2.1 Data Source and Description

The `aids` dataset is structured as a data frame, sourced from the `JM` package in `R` (Rizopoulos (2010)), comprising 1408 observations across 9 key variables. These variables provide a detailed account of patient information, treatment response, and clinical outcomes, as follows: (1). Patient Identifier (`patient`): This variable serves as a unique identifier for each patient. The dataset encompasses data from a total of 467 different patients, allowing for a diverse and extensive analysis. (2) Time to Event (`Time`): This variable records the time to either death or censoring, critical for survival analysis. (3) Event Indicator (`death`): A numeric vector where '0' denotes censoring, and '1' indicates the occurrence of death. (4). CD4 Cell

Count (`CD4`): This is a key biomarker in HIV treatment, with each entry representing the count of CD4 cells. (5). Observation Time (`obstime`): This records the specific time points at which the CD4 cell count was measured, crucial for longitudinal analysis. (6). Drug Type (`drug`): A categorical variable indicating the type of antiretroviral drug administered - ddC for Zalcitabine and ddI for Didanosine. (7). Gender (`gender`): This factor distinguishes between male and female patients. (8). Previous Opportunistic Infection (`prevOI`): Categorized as 'AIDS' for patients with a previous opportunistic infection (or AIDS diagnosis) at study entry, and 'noAIDS' for those without. (9). AZT Response (`AZT`): Indicates the patient's response to prior AZT treatment, classified as either 'intolerance' or 'failure'. Sample sizes and descriptive statistics for key variables at study entry are shown in Table 1.

Table 1: Descriptive statistics of baseline characteristics of the 467 subjects by treatment groups.

	ddC	ddI	Overall
	(N=237)	(N=230)	(N=467)
Gender			
Male	214 (90.3%)	208 (90.4%)	422 (90.4%)
Female	23 (9.7%)	22 (9.6%)	45 (9.6%)
Previous AIDS diagnosis			
No	79 (33.3%)	81 (35.2%)	160 (34.3%)
Yes	158 (66.7%)	149 (64.8%)	307 (65.7%)
AZT therapy			
Intolerance	146 (61.6%)	146 (63.5%)	292 (62.5%)
Failure	91 (38.4%)	84 (36.5%)	175 (37.5%)
Baseline CD4 cells count			
Mean (SD)	7.02 (4.65)	7.24 (4.78)	7.13 (4.71)
Median [Min, Max]	5.74 [0, 19.1]	6.32 [0, 19.2]	6.08 [0, 19.2]

It is important to note that the `aids.id` data frame, a subset of the main dataset, includes the initial CD4 cell count measurement for each patient. This subset is particularly utilized for fitting the survival model, offering an initial snapshot of each patient's immune status.

2.2 Joint Modeling Framework

For our study's context, we will set up a joint model (JM) that incorporates both a linear mixed effects model (LME) for the longitudinal outcome of CD4 cell count and a survival model for the time-to-event data.

Consider Y_{ij} to be the CD4 cell count for the i-th patient at the j-th time point. We structure the LME for the longitudinal CD4 measurements as follows:

$$Y_{ij} = \beta_0 + \beta_1 \texttt{obstime}_{ij} + \beta_2 \texttt{drug}_i + \beta_3 \texttt{gender}_i + \beta_4 \texttt{prevOI}_i + \beta_5 \texttt{AZT}_i + b_{0i} + b_{1i} \texttt{obstime}_{ij} + \epsilon_{ij},$$

where β_i's are the fixed effects, b_{0i} and b_{1i} are random intercept and slope terms that allow for patient-specific deviations from the population average trajectory, and ϵ_{ij}'s are the independent residual errors. The random effects are assumed to follow a joint-normal distribution with mean zero and are uncorrelated with the residual errors.

Having specified the LME, we now consider the survival submodel of the JM. As shown in the previous literatures (Mellors et al. (1996), Volberding et al. (1990)), the absolute number of CD4 lymphocyte cells in the peripheral blood is used extensively as a prognostic factor and a surrogate marker for progression of disease and for death in clinical studies of human immunodeficiency virus (HIV) infection.

$$h_i(t) = h_0(t) \exp(\gamma_0 + \gamma_1 \texttt{drug}_i + \alpha m_i(t)),$$

where $h_0(t)$ represents the baseline hazard function, parameterized by a Weibull distribution in our case, and $m_i(t)$ is the true but unobserved value of CD4 for the i-th patient at time

t, estimated by the LME model we introduced above. The γ coefficients correspond to the effects of the baseline covariates such as gender, previous opportunistic infection status (`prevOI`), AZT response, and the specific antiretroviral treatment (`drug`) and it interaction with `prevOI`. The coefficient α_1 quantifies the association between the true CD4 count and the hazard of the event, controlling for the other covariates in the model.

This JM allows us to assess the direct impact of the covariates on the hazard function while controlling for the CD4 count's progression over time. The shared random effects in the longitudinal submodel are assumed to correlate with the survival outcomes, thereby capturing the joint distribution of the longitudinal and survival data. This integrated approach provides a deeper understanding of how baseline characteristics and the evolution of CD4 counts influence patient survival.

Estimation and inference within the joint modeling framework we are employing are predicated on key conditional independence assumptions. Specifically, these assumptions underlie the submodels detailed earlier—namely the longitudinal model for CD4 counts and the survival model for time-to-event data. These assumptions posit that, given the random effects, the longitudinal and survival processes are independent. This extends to the independence of the repeated longitudinal measures within the CD4 submodel.

Moreover, we adopt the assumption that the right-censoring mechanism and the scheduling of the longitudinal observations occur independently of the actual event times and future CD4 measurements. This is critical to ensure unbiased estimates and valid inferences within our analysis.

2.3 Estimation and Inference

To maximize the log-likelihood of our joint model, we utilize standard estimation algorithms. These procedures are well-established within the statistical community and are implemented in various software packages like `joineRML` from Hickey et al. (2018). For our analysis, we will be utilizing the `JM` package in `R` (Rizopoulos (2010)), which is specifically designed for the estimation of joint models. This package provides the necessary tools to efficiently carry out the estimation process and to draw rigorous inferences from our joint modeling analysis.

2.4 Expected Survival

Once the JM was estimated using maximum likelihood estimation, we computed predicted survival probabilities for each individual in our study. Participants had longitudinal data collected from the time of their treatment initiation up to—but not including—the time of their event or censoring. The last known time point without an event was denoted as t, and at this time, the survival probability was set to 1, as it was the last observed time of no risk for the event. If a participant was censored, then t represented the age of censoring.

Our interest centered on the probability of survival beyond a certain time after t, denoted as u, conditional on the data and covariates. The notation $\pi_i(u|t)$ indicates the i-th individual's probability of not experiencing the event by time u, given they had not experienced it by time t.

In our study, the estimation of $\pi_i(u|t)$ was carried out using a Markov Chain Monte Carlo (MCMC) methodology. This estimation was performed subsequent to fitting our joint model, which combines longitudinal and time-to-event data. The computation of $\pi_i(u|t)$ was conducted using the `survfitJM` function from the `JM` package (details see Rizopoulos (2010)),

which interfaces with the MCMC output. For each iteration within the MCMC sequence, a single draw from the posterior distribution of the model parameters, including the random effects, was taken. These parameters were then applied to calculate the survival probability for an individual at each time point beyond t.

This procedure was replicated across 500 MCMC iterations to form a distribution of predicted survival probabilities for each individual at age u. The mean of these probabilities across all iterations was considered as the predicted survival probability for that individual. The standard deviation of the probabilities provided an estimate of uncertainty, which was used to compute the 95% confidence intervals for the survival curves.

By employing the MCMC method, we not only accounted for the uncertainty in the model parameters but also incorporated the random effects variation, resulting in a robust probabilistic framework for our survival probability estimates.

2.5 Comparison with Traditional Survival Analysis with Time-dependent Variable

To evaluate the efficacy of the joint modeling approach relative to traditional methods, we will conduct a comparative analysis against the time-dependent Cox proportional hazards model. The traditional Cox model will be specified with the drug and CD4 count as a time-varying covariate, that is

$$h_i(t) = h_0(t) \exp(\gamma_1 \texttt{drug}_i + \theta \texttt{CD4}_i(t)),$$

we will use the `coxph` function from the `survival` package (Therneau (2023)) in `R`. Upon fitting both models to the data, we will compare them based on several criteria, including

the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood values, to assess model fit and complexity. We will also compare the regression coefficients results from both models to see which model gives us a better understanding on the CD4 cell counts and the treatment effects on the survival outcome.

3 Results

3.1 Longitudinal and Survival Outcomes for the Patients

The study was designed to follow patients over an 18-month period, with CD4 cell counts measured at multiple time points: at the inception of the study where randomization occurred, and subsequently at 2, 6, 12, and 18 months post-randomization. These CD4 cell counts serve as critical biomarkers for monitoring the progression of HIV and the response to treatment. CD4 cell counts typically demonstrate a distribution that is right-skewed. To address this skewness and facilitate a more robust analysis, we will proceed with the square root transformation of the CD4 cell count values. Figure 1 shows the patients evolutions in time of the square root of the CD4 cell counts stratified by the baseline variables at the randomization. We note that patients in both treatment groups exhibit comparable levels of variability in their longitudinal profiles, however, patients longitudinal profiles are different in other three baseline groups.

For the patients survival, mong all 237 patients in `ddC` treatment group, 88 (37.1%) patients had died, where in `ddI` group, 100 out of 230 (43.5%) patients had died. Figure 2 shows the Kaplan-Meier estimate of the survival functions between two treatment groups. From the plot we can see that the `ddC` group has slightly higher survival than the `ddI` group after the six month of follow-up.

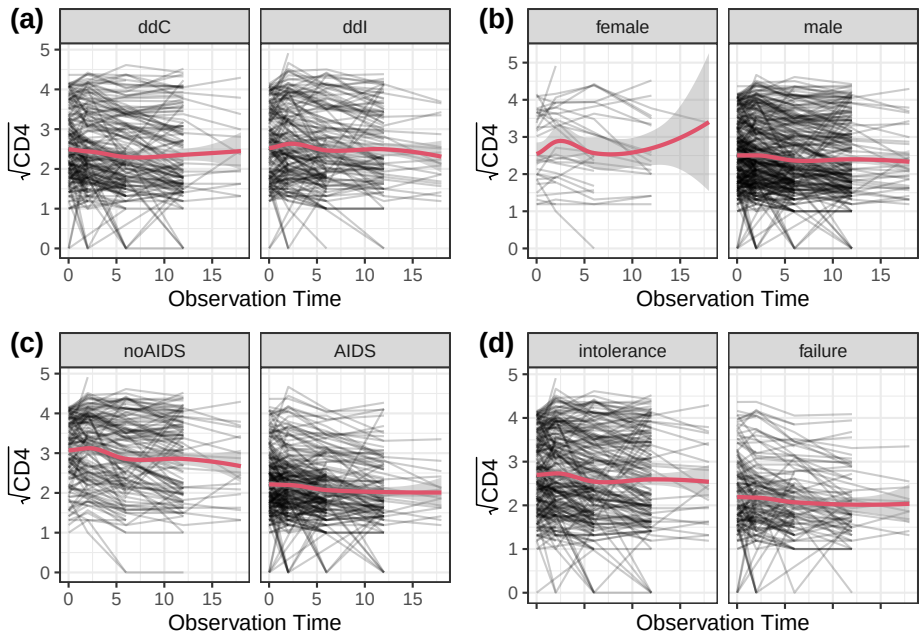

Figure 1: Patients evolutions in time of the square root of the CD4 cell counts. (a) Stratified by `drug`. (b) Stratified by `gender`. (c) Stratified by `prevOI`. (d) Stratified by `AZT`.

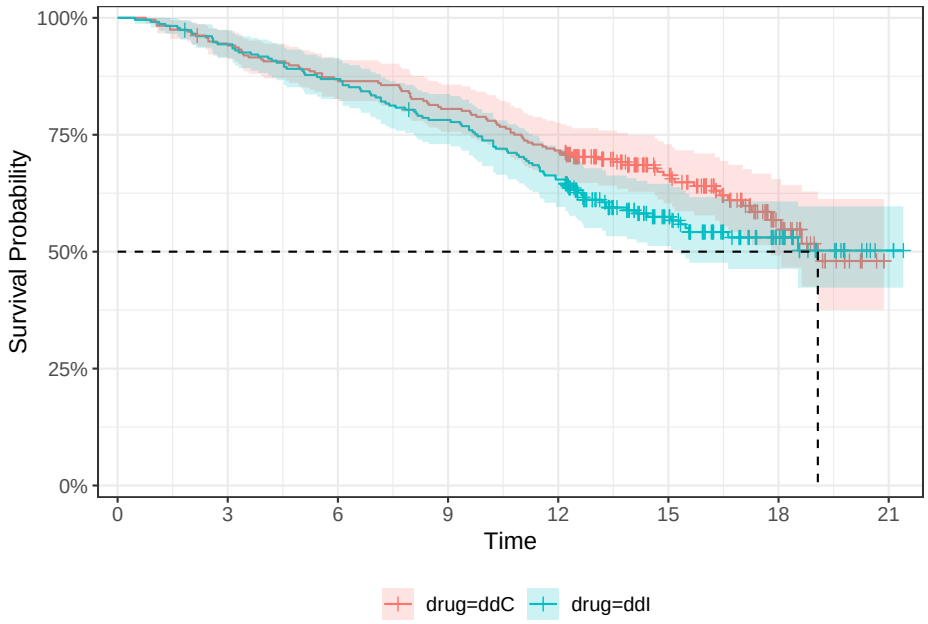

Figure 2: Kaplan-Meier estimates of the survival function for the patient in different treatment groups.

3.2 Comparison of Coefficients for Joint Model and Cox Proportional Hazard Model

Table 2 shows the coefficients estimates from a naive Cox proportional hazard model with time-dependent parameter $\sqrt{\text{CD4}}$ and the treatment, and from the joint model. From the results we can see that after incorporating the square root of the CD4 count as a covariate in our Cox model, the results did not provide substantial evidence of a treatment effect (HR(95% CI): 1.17 (0.88-1.56)). Building on this, we moved to develop and implement a joint model. The longitudinal submodel and survival submodel specifications are stated before. For the survival submodel of the joint model, mirroring the approach taken in the Cox model, we included the treatment effect as a time-independent covariate. Additionally, we incorporated the estimated true effect of the CD4 cell count, as derived from the longitudinal model, as a time-varying covariate.

Table 2: Parameter estimates for the multivariate joint model and Cox proportional hazard model with time-dependent covariate.

Joint Model				Cox Proportional Hazard Model			
$\text{Variables}_{\text{JM}}$	$\text{Estimates}_{\text{JM}}$	SD_{JM}	$\text{p-value}_{\text{JM}}$	$\text{Variables}_{\text{Cox}}$	$\text{Estimates}_{\text{Cox}}$	SD_{Cox}	$\text{p-value}_{\text{Cox}}$
Event Model							
Intercept	-2.164	0.345	< 0.001				
`drug = ddI`	0.340	0.152	0.025	`drug = ddI`	0.158	0.146	1.081
α^*	-1.103	0.120	< 0.001	$\sqrt{\texttt{CD4}}$	-0.544	0.075	-7.241
$\log(\lambda)^\dagger$	0.239	0.072	< 0.001				
Longitudinal Model							
Intercept	3.086	0.134	< 0.001				
`obstime`	-0.038	0.003	< 0.001				
`drug = ddI`	0.069	0.074	0.352				
`gender = male`	0.000	0.127	0.999				
`prevOI = AIDS`	-0.882	0.094	< 0.001				
`AZT = failure`	-0.074	0.092	0.42				

* The effect of the true square root CD4 cell count in the risk for death.

† The logarithm of the estimated shape parameter of Weibull distribution.

The comparison of the standard time-dependent Cox model with our joint model uncovers

several interesting aspects. Specifically, the coefficient for `ddI` in the joint model is more substantial in magnitude compared to the Cox model, and it is significant (HR(95% CI): 1.40(1.04-1.89)). This difference suggests an enhanced treatment effect in the joint model analysis. Additionally, a more pronounced discrepancy is evident in the estimated effects of the CD4 cell count. The time-dependent Cox model yields a regression coefficient of -0.54, while the joint model estimates this coefficient at -1.34 (the α in JM), indicating a more substantial influence of CD4 count in the joint model framework.

3.3 Expected Patient Survival

Figure 3 shows the Monte Carlo estimates of two patients survivals, $\pi_i(u \mid t)$, at two different time points, with the history of longitudinal outcomes up to t. Notice that the estimated $\pi_i(u \mid t)$ for each subject at each time points is calculated using the history of longitudinal outcomes that are later than the time of the last available observation. This approach is adopted because, for earlier time points preceding the last observed longitudinal outcome, it is already established that the subjects were alive and therefore $\pi_i(u \mid t) = 1$.

From the plot we can see that, Patient 12 has CD4 cell count measurements up to 18 months from randomization and he was lost to follow-up after 18.1 months. Using only 2 longitudinal outcomes (upper left panel), he has a estimated 78.9% probability of surviving more than 21.5 months (95% CI: 43.5% - 93.4%), while using 4 longitudinal outcomes (upper right panel), he has a estimated 84.1% probability of surviving more than 21.5 months (95% CI: 69.6% - 93.8%), which is larger than the estimate using 2 longitudinal outcomes, and it has a narrower CI. Same as to Patient 157. We get this more accurate estimates since JM explicitly accounts for measurement errors in the longitudinal process, which can lead to more accurate estimates of the association between the CD4 count and the risk of the death.

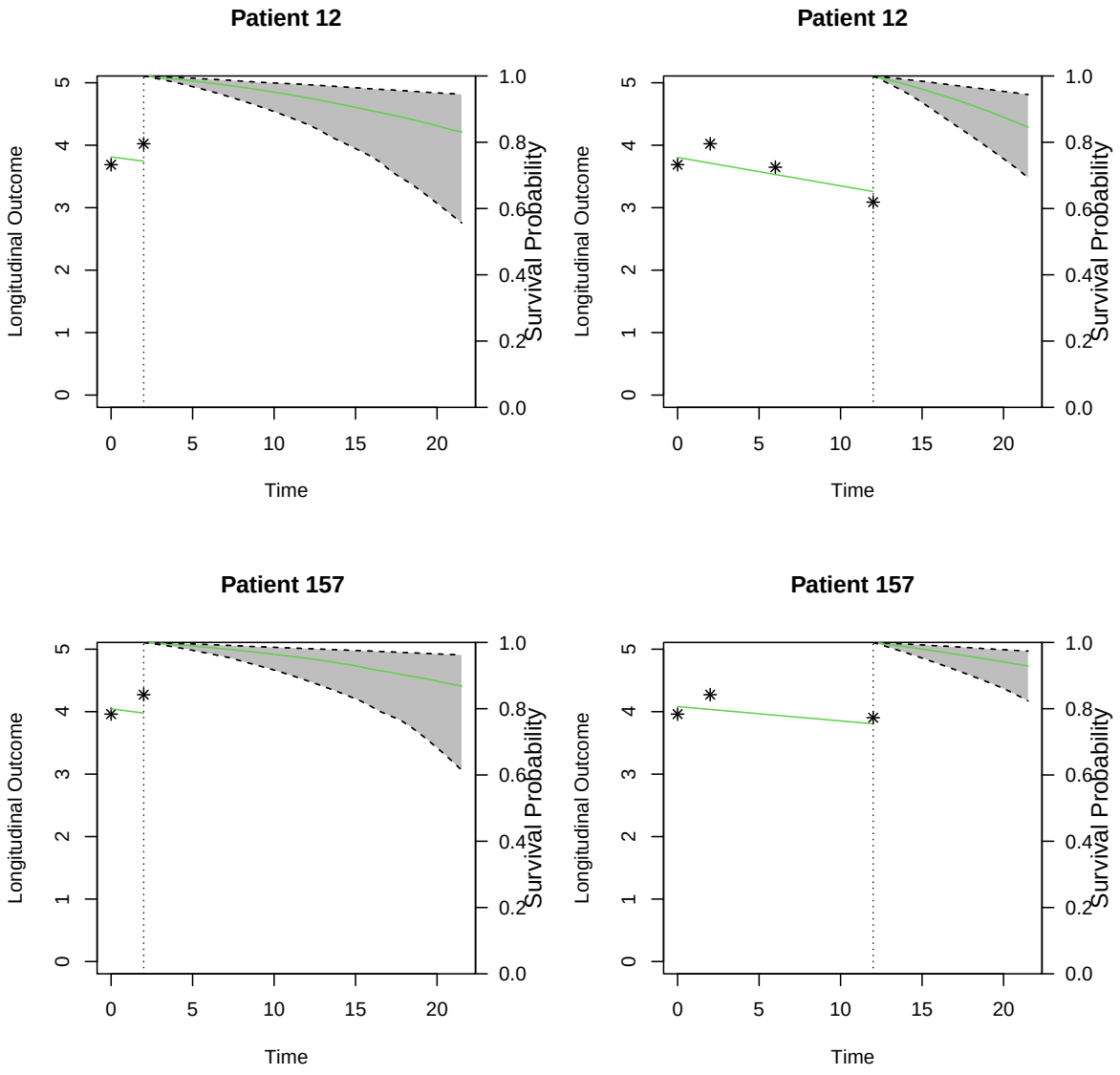

Figure 3: Predicted survival probabilities for Patient 12 (upper left panel) and Patient 157 (lower left panel) including available longitudinal $\sqrt{CD4}$ cell counts measurements (upper right and lower right panels), based on 200 Monte Carlo samples. The green solid line depicts the median of $\pi_i(u \mid t)$ over the Monte Carlo samples. The black dashed lines depict a 95% pointwise confidence interval.

4 Discussion

Our analysis comparing Joint Modeling (JM) with the time-dependent Cox model in a longitudinal HIV study highlights the robustness of JM in capturing the complex interplay between treatment effects and CD4 cell count trajectories. Notably, JM revealed a statistically significant and larger treatment effect compared to the Cox model, emphasizing its capacity to integrate longitudinal biomarker data with survival outcomes. This aligns well with personalized medicine approaches, where understanding individual patient trajectories is key.

Dynamic predictions using JM, based on longitudinal outcomes, demonstrated enhanced precision in survival estimates. This suggests that continuous monitoring of biomarkers like CD4 counts can significantly refine survival predictions, offering valuable insights for clinical decision-making in HIV treatment.

However, JM comes with its own set of limitations. One of the primary challenges is the computational complexity and increased demands on data structure and quality. JM requires careful handling of missing data and accurate modeling of the random effects structure, which can be computationally intensive. Moreover, the assumptions underlying the joint model, such as the relationship between the longitudinal and survival processes, need careful consideration as they can impact the model's interpretation and conclusions.

Despite these limitations, our findings endorse the use of JM in clinical research, particularly in scenarios involving longitudinal and time-to-event data. Future research should focus on addressing the computational challenges and exploring more flexible models that can handle complex data scenarios commonly encountered in medical research.

Appendix

A.1 Model Diagnostics for Joint Modelling

Table A.1: Comparing model fits between time-dependent Cox model and JM.

Models	log-likelihood	DF	AIC	BIC
Time-dependent Cox	-994.078	2	1992.157	1998.630
JM	-2039.605	14	4107.211	4165.259

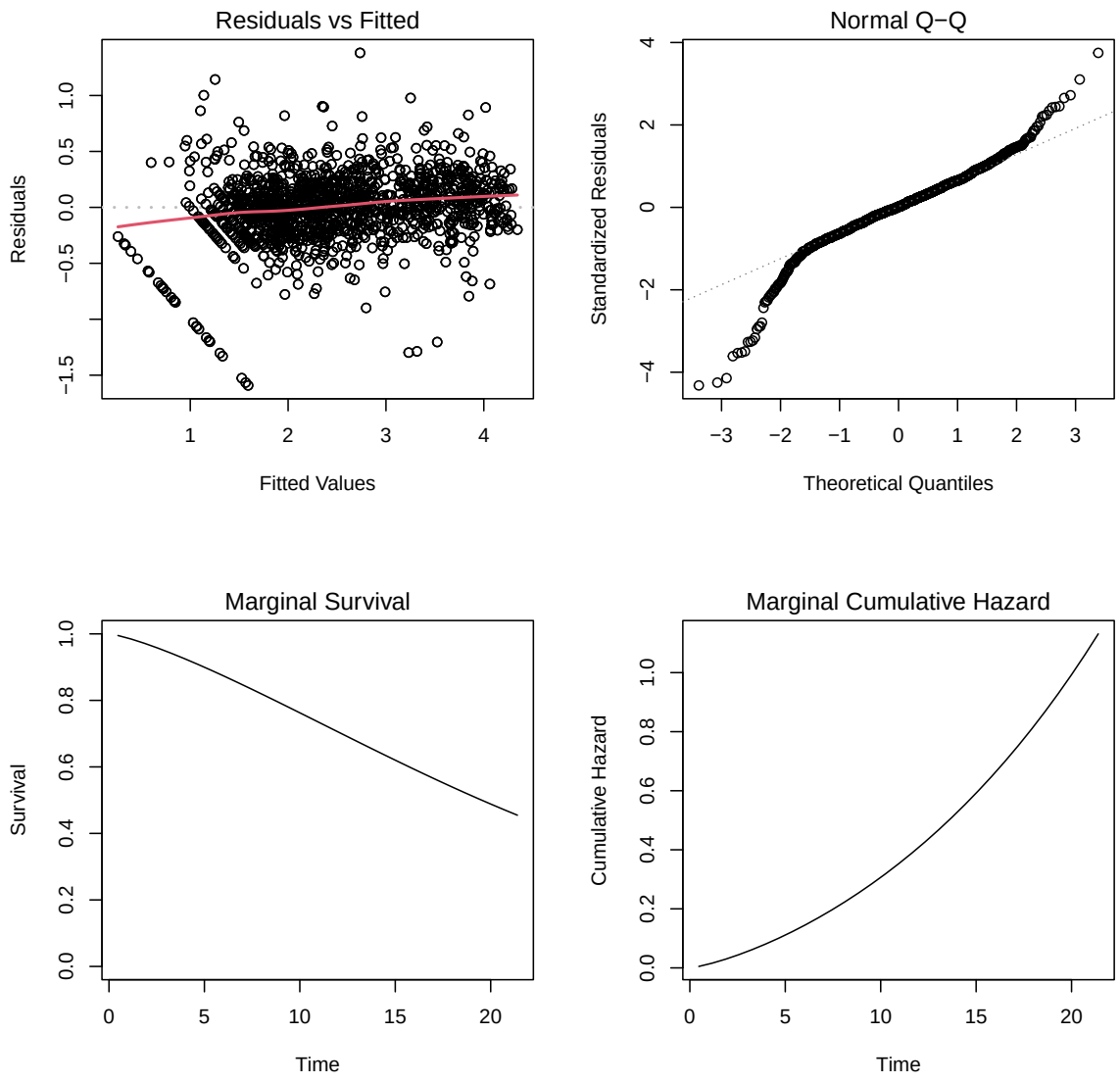

Figure A.1: Diagnostic plots for the fitted joint model. Upper left panel: subject specific residuals for the longitudinal process versus their corresponding fitted values. Upper right panel: normal Q-Q plot of the standardized subject specific residuals for the longitudinal process. Lower left panel: an estimate of the marginal survival function for the event process. Lower right panel: an estimate of the marginal cumulative risk function for the event process.

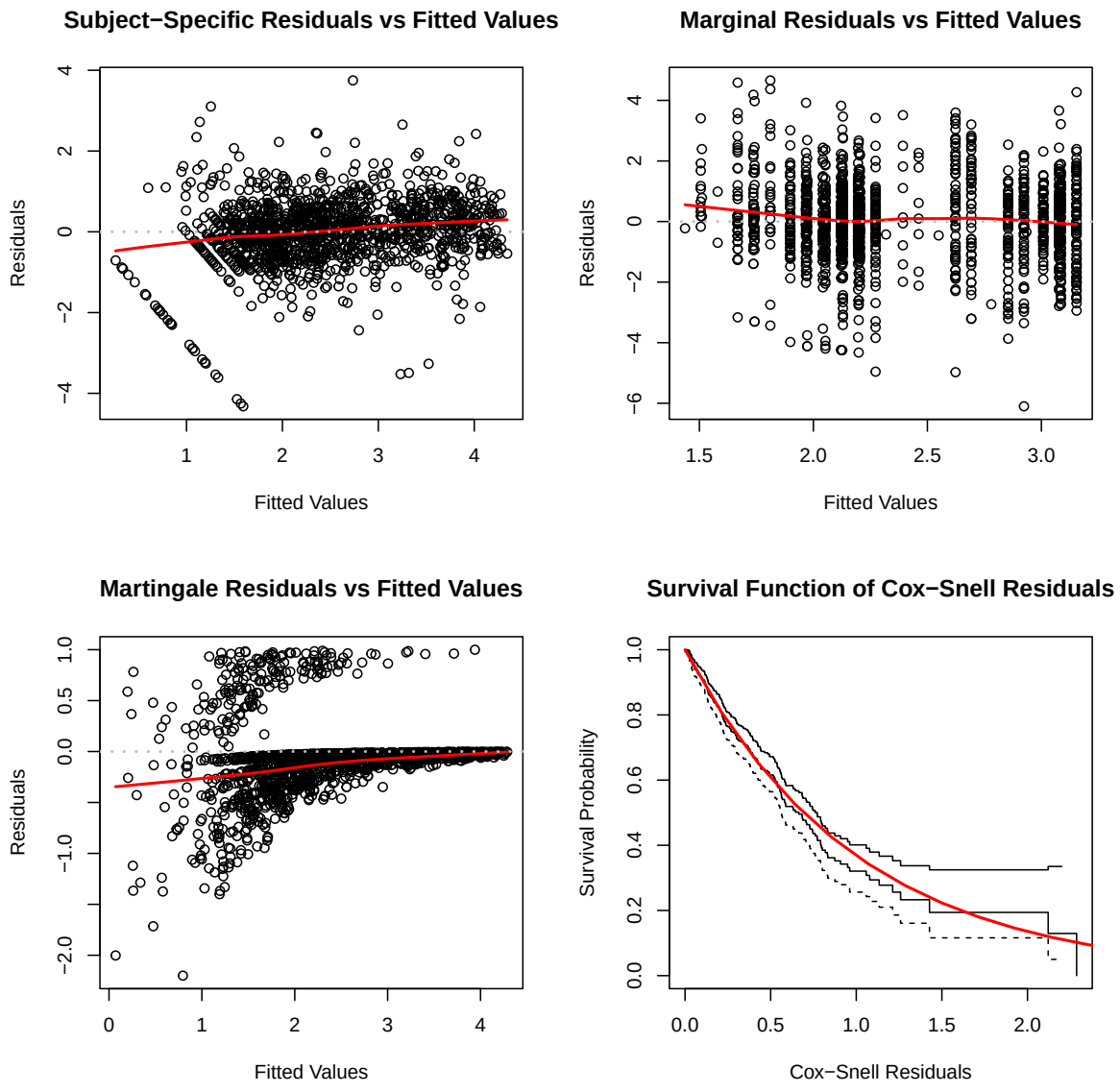

Figure A.2: Diagnostic plots for the fitted joint model. The dashed lines in lower right panel denote the 95% confidence intervals for the Kaplan-Meier estimate of the Cox-Snell residuals.

A.2 Code for This Report

```
knitr::opts_chunk$set(echo = FALSE, message = F, warning = F, out.width = "75%", fig.align = "center")
options(knitr.kable.NA = '')
library(tidyverse)
library(caret)
library(latex2exp)
library(gstat)
library(sp)
library(nlme)
library(knitr)
library(kableExtra)
library(geepack)
library(survival)
library(JM)
library(survminer)
library(ggsurvfit)
library(lme4)
write_matex <- function(x) {
  begin <- "$$\\begin{bmatrix}"
  end <- "\\end{bmatrix}$$"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  writeLines(c(begin, X, end))
}
theme_set(
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
```

```
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )
)
aids.data <- JM::aids
aids.surv.data <- JM::aids.id
library(table1)
#names(aids.surv.data)
aids.tbl.data <- aids.surv.data
aids.tbl.data$gender <- factor(aids.tbl.data$gender, levels = c("male", "female"), labels = c("Male", "
aids.tbl.data$prevOI <- factor(aids.tbl.data$prevOI, levels = c("noAIDS", "AIDS"), labels = c("No", "Yes
aids.tbl.data$AZT <- factor(aids.tbl.data$AZT, levels = c("intolerance", "failure"), labels = c("Intole
aids.tbl.data$drug <- factor(aids.tbl.data$drug, levels = c("ddC", "ddI"), labels = c("ddC", "ddI"))

label(aids.tbl.data$gender) = "Gender"
label(aids.tbl.data$prevOI) = "Previous AIDS diagnosis"
label(aids.tbl.data$AZT) = "AZT therapy"
label(aids.tbl.data$drug) = "Treatment"
label(aids.tbl.data$CD4) = "Baseline CD4 cells count"

pvalue <- function(x, ...) {
    # Construct vectors of data y, and groups (strata) g
    y <- unlist(x)
    g <- factor(rep(1:length(x), times=sapply(x, length)))
    if (is.numeric(y)) {
        # For numeric variables, perform a standard 2-sample t-test
        p <- t.test(y ~ g)$p.value
    } else {
        # For categorical variables, perform a chi-squared test of independence
        p <- chisq.test(table(y, g))$p.value
    }
    # Format the p-value, using an HTML entity for the less-than sign.
    # The initial empty string places the output on the line below the variable label.
    c("", sub("<", "&lt;", format.pval(p, digits=3, eps=0.001)))
}
```

```r
table1(~ gender + prevOI + AZT + CD4 | drug,
    data=aids.tbl.data, caption="Descriptive statistics of baseline characteristics of the 467 subjects
# ggplot(aids.data, aes(x = obstime, y = sqrt(CD4))) +
#     geom_line(aes(group = patient), alpha = 0.2) +
#     labs(title = "Spaghetti Plot of CD4 by Time (Stratified by Drug)",
#          x = "Observation Time",
#          y = "CD4 Count") +
#   geom_smooth(method = "loess", formula = y ~ x,se = TRUE, aes(linetype = "LOESS", color = "LOESS"),
#   theme_bw() +
#   scale_linetype_manual(name = "Method", values = c(1),breaks = c( "LOESS"))+
#   scale_color_manual(name = "Method", values = c(2),breaks = c( "LOESS")) +
#   theme(legend.position = "none")
p.drug <-
ggplot(aids.data, aes(x = obstime, y = sqrt(CD4))) +
    geom_line(aes(group = patient), alpha = 0.2) +
    facet_wrap(~ drug) +
    labs(#title = TeX("Spaghetti Plot of $\\sqrt{CD4}$ by Time (Stratified by drug)"),
         x = "Observation Time",
         y = TeX("$\\sqrt{CD4}$")) +
  geom_smooth(method = "loess", formula = y ~ x,se = TRUE, aes(linetype = "LOESS", color = "LOESS"), li
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c( "LOESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "LOESS")) +
  theme(legend.position = "none")

p.gender <-
ggplot(aids.data, aes(x = obstime, y = sqrt(CD4))) +
    geom_line(aes(group = patient), alpha = 0.2) +
    facet_wrap(~ gender) +
    labs(#title = "Spaghetti Plot of CD4 by Time (Stratified by Gender)",
         x = "Observation Time",
         y = TeX("$\\sqrt{CD4}$")) +
  geom_smooth(method = "loess", formula = y ~ x,se = TRUE, aes(linetype = "LOESS", color = "LOESS"), li
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c( "LOESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "LOESS")) +
```

```
  theme(legend.position = "none")

p.prevOI <-
ggplot(aids.data, aes(x = obstime, y = sqrt(CD4))) +
    geom_line(aes(group = patient), alpha = 0.2) +
    facet_wrap(~ prevOI) +
    labs(#title = "Spaghetti Plot of CD4 by Time (Stratified by Previous AIDS Diagnosis)",
         x = "Observation Time",
         y = TeX("$\\sqrt{CD4}$")) +
  geom_smooth(method = "loess", formula = y ~ x,se = TRUE, aes(linetype = "LOESS", color = "LOESS"), li
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c( "LOESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "LOESS")) +
  theme(legend.position = "none")

p.AZT <-
ggplot(aids.data, aes(x = obstime, y = sqrt(CD4))) +
    geom_line(aes(group = patient), alpha = 0.2) +
    facet_wrap(~ AZT) +
    labs(#title = "Spaghetti Plot of CD4 by Time (Stratified by AZT Therapy)",
         x = "Observation Time",
         y = TeX("$\\sqrt{CD4}$")) +
  geom_smooth(method = "loess", formula = y ~ x,se = TRUE, aes(linetype = "LOESS", color = "LOESS"), li
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c( "LOESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "LOESS")) +
  theme(legend.position = "none")


library(ggpubr)

ggarrange(p.drug , p.gender, p.prevOI, p.AZT + rremove("x.text"),
          labels = c("(a)", "(b)", "(c)", "(d)"),
          ncol = 2, nrow = 2)
km.fit <- survfit(Surv(Time, death) ~ drug, aids.surv.data)
km.fit %>% ggsurvfit() + add_censor_mark() + add_confidence_interval() + add_quantile() + scale_ggsurvfi
```

```r
# Different models for longitudinal process
# "CD4" "obstime" "drug" "gender"  "prevOI"  "AZT"
lme.null <- lme(sqrt(CD4) ~ obstime, random = ~ 1 | patient, data = aids.data)
#summary(lme.null)
lme.simple <- lme(sqrt(CD4) ~ obstime + drug + gender + prevOI + AZT, random = ~ 1 | patient, data = aid
#summary(lme.simple)
lme.complex<- lme(sqrt(CD4) ~ obstime + I(obstime^2) + drug + gender + prevOI + AZT + drug*(obstime + pr
#summary(lme.complex)
lme.complex<- lme(sqrt(CD4) ~ obstime + drug + gender + prevOI + AZT + drug*(obstime + prevOI + AZT + ge
#summary(lme.complex)
#anova.lme(lme.null, lme.simple, lme.complex) # lme.simple has the smallest AIC

lme.simple.slope <- lme(sqrt(CD4) ~ obstime + drug + gender + prevOI + AZT, reStruct(~ obstime | patient
#summary(lme.simple.slope)
lme.simple.slope2 <- lme(sqrt(CD4) ~ obstime + drug + gender + prevOI + AZT, reStruct(~ obstime + prevOI



lme.simple.slope.ns <- lme(sqrt(CD4) ~ ns(obstime,1) + drug + gender + prevOI + AZT, reStruct(~ obstime
#anova.lme(lme.simple,lme.simple.slope, lme.simple.slope2)
cox.null <- coxph(Surv(Time, death) ~ 1, data = aids.surv.data, x = TRUE,y = TRUE,model = TRUE)
#summary(cox.null)
cox.simple <- coxph(Surv(Time, death) ~ drug, data = aids.surv.data, x = TRUE,y = TRUE,model = TRUE)
#summary(cox.simple)
cox.complex <- coxph(Surv(Time, death) ~ drug + gender + prevOI + AZT + drug*prevOI, data = aids.surv.da
#summary(cox.complex)
#library(JMbayes)
JM1 <- jointModel(lme.simple.slope, cox.complex, timeVar = "obstime", method = "weibull-PH-aGH")
sum.JM.1 <- summary(JM1)
JM2 <- jointModel(lme.simple.slope, cox.simple, timeVar = "obstime", method = "weibull-PH-aGH")
sum.JM.2 <-summary(JM2)

#plot(JM::survfitJM(JM2, newdata = aids[aids$patient == "7", ], idVar = "patient", M = 50))
cox.td <- coxph(Surv(start, stop, death) ~ drug + sqrt(CD4), data = aids.surv.data, x = TRUE,y = TRUE,m
sum.cox.td <- summary(cox.td)
# Extract coefficients from the Joint Model
JM_Coef_Event <- as.data.frame(sum.JM.2$`CoefTable-Event`)
```

```r
JM_Coef_Long <- as.data.frame(sum.JM.2$`CoefTable-Long`)

# Combine Joint Model coefficients into one data frame
JM_Coef <- rbind(JM_Coef_Event, JM_Coef_Long) %>% mutate(`p-value` = ifelse(`p-value` < 0.001, "< 0.001

# Extract coefficients from Cox model
Cox_Coef <- as.data.frame(coef(summary(cox.td)))
Cox_Coef <- Cox_Coef %>% dplyr::select(-`exp(coef)`)

colnames(Cox_Coef) <- colnames(JM_Coef)

JM_Coef$param <- row.names(JM_Coef)
Cox_Coef$param <- row.names(Cox_Coef)
Cox_Coef_null <- as.data.frame(matrix(rep(NA, 35), byrow = T, nrow = 7))
colnames(Cox_Coef_null) <- colnames(Cox_Coef)
Cox_Coef_line <- as.data.frame(matrix(rep(NA, 5), byrow = T, nrow = 1))
colnames(Cox_Coef_line) <- colnames(Cox_Coef)
Cox_Coef <- rbind(Cox_Coef_line,Cox_Coef, Cox_Coef_null) %>% mutate(`p-value` = ifelse(`p-value` < 0.00

coef.tab <- cbind(JM_Coef %>% dplyr::select(param, everything()), Cox_Coef %>% dplyr::select(param, eve
coef.tab$param <- c(
  "Intercept",
  "$\\texttt{drug = ddI}$",
  "$\\alpha$",
  "$\\log(\\lambda)$",
  "Intercept",
  "$\\texttt{obstime}$",
  "$\\texttt{drug = ddI}$",
  "$\\texttt{gender = male}$",
  "$\\texttt{prevOI = AIDS}$",
  "$\\texttt{AZT = failure}$"
)

coef.tab$param[3]<- paste0(coef.tab$param[3],footnote_marker_symbol(1))
coef.tab$param[4]<- paste0(coef.tab$param[4],footnote_marker_symbol(2))
```

```r
coef.tab[,6] <- c(NA,"$\\texttt{drug = ddI}$", "$\\sqrt{\\texttt{CD4}}$", NA, NA, NA, NA, NA, NA, NA)

coef.tab[, -4][,-9] %>%  kable(format = "latex",digits = 3, booktab = T, row.names = F, col.names = c("
  footnote(symbol = c("The effect of the true square root CD4 cell count in the risk for death.", "The
#plot(JM::survfitJM(JM2, newdata = aids[aids$patient == "11", ], idVar = "patient", M = 50))

#plot(JM::survfitJM(JM2, newdata = aids[aids$patient == "157", ], idVar = "patient", M = 50))

# sample patients
aids.data.12 <- aids.data[aids.data$patient == "12", ]
#nrow(aids.data.12) #5
aids.data.157 <- aids.data[aids.data$patient == "157", ]
# 4
# plot the data
sfit.12.1<- survfitJM(JM2, newdata = aids.data.12[1:2, ], idVar = "patient")
sfit.12.2 <- survfitJM(JM2, newdata = aids.data.12[1:4, ], idVar = "patient")

sfit.157.1<- survfitJM(JM2, newdata = aids.data.157[1:2, ], idVar = "patient")
sfit.157.2 <- survfitJM(JM2, newdata = aids.data.157[1:3, ], idVar = "patient")
par(mfrow=c(2,2))
plotfit.12.1 <- plot(sfit.12.1, estimator="median", include.y = TRUE, conf.int=0.95, fill.area=TRUE, co
plotfit.12.2 <- plot(sfit.12.2, estimator="median", include.y = TRUE, conf.int=0.95, fill.area=TRUE, co
plotfit.157.1 <- plot(sfit.157.1, estimator="median", include.y = TRUE, conf.int=0.95, fill.area=TRUE,
plotfit.157.2 <- plot(sfit.157.2, estimator="median", include.y = TRUE, conf.int=0.95, fill.area=TRUE,
cat('\\setcounter{figure}{0}')  # Reset figure counter
cat('\\renewcommand{\\thefigure}{A.\\arabic{figure}}')  # Change format to A.x

par(mfrow = c(2,2))
plot(JM2)
plotResid <- function (x, y, ...) {
plot(x, y, ...)
lines(lowess(x, y), col = "red", lwd = 2)
abline(h = 0, lty = 3, col = "grey", lwd = 2)
}
par(mfrow = c(2, 2))
```

```r
resSubY <- residuals(JM2, process = "Longitudinal", type = "stand-Subject")
fitSubY <- fitted(JM2, process = "Longitudinal", type = "Subject")
plotResid(fitSubY, resSubY, xlab = "Fitted Values", ylab = "Residuals",main = "Subject-Specific Residual
resMargY <- residuals(JM2, process = "Longitudinal", type = "stand-Marginal")
fitMargY <- fitted(JM2, process = "Longitudinal", type = "Marginal")
plotResid(fitMargY, resMargY, xlab = "Fitted Values", ylab = "Residuals", main = "Marginal Residuals vs
resMartT <- residuals(JM2, process = "Event", type = "Martingale")
fitSubY <- fitted(JM2, process = "Longitudinal", type = "EventTime")
plotResid(fitSubY, resMartT, xlab = "Fitted Values", ylab = "Residuals", main = "Martingale Residuals vs
resCST <- residuals(JM2, process = "Event", type = "CoxSnell")
sfit <- survfit(Surv(resCST, death) ~ 1, data = aids.surv.data)
plot(sfit, mark.time = FALSE, conf.int = TRUE, lty = 1:2,  xlab = "Cox-Snell Residuals", ylab = "Survival
curve(exp(-x), from = 0, to = max(aids.surv.data$Time), add = TRUE, col = "red", lwd = 2)

cat('\\setcounter{table}{0}')  # Reset table counter
cat('\\renewcommand{\\thetable}{A.\\arabic{table}}')  # Change format to A.x
# AIC(cox.td)
# BIC(cox.td)
# logLik(cox.td)
# AIC(JM2)
# logLik(JM2)

model.fit.tab <- tibble(
  models = c("Time-dependent Cox", "JM"),
  log.lik = c(logLik(cox.td), logLik(JM2)),
  d.f = c(2,14),
  AIC = c(AIC(cox.td), AIC(JM2)),
  BIC = c(BIC(cox.td), BIC(JM2))
)
model.fit.tab %>% kable(format = "latex",digits = 3, booktab = T, row.names = F, col.names = c("Models"
```

References

Goldman, A. I., Carlin, B. P., Crane, L. R., Launer, C., Korvick, J. A., Deyton, L., and Abrams, D. I. (1996). Response of cd4 lymphocytes and clinical consequences of treatment using ddi or ddc in patients with

advanced hiv infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 11(2):161–169.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Medical Research Methodology*, 18(1).

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796–2801.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Little, R. and Rubin, D. (2019). *Statistical Analysis with Missing Data, Third Edition*. Wiley.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., and Kingsley, L. A. (1996). Prognosis in hiv-1 infection predicted by the quantity of virus in plasma. *Science*, 272(5265):1167–1170.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Wiley.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33.

Rizopoulos, D. and Lesaffre, E. (2014). Introduction to the special issue on joint modelling techniques. *Statistical Methods in Medical Research*, 23(1):3–10.

Therneau, T. M. (2023). *A Package for Survival Analysis in R*. R package version 3.5-7.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer New York.

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

Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, W. D., Soeiro, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hyslop, N. E., Richman, D. D., Valentine, F. T., and Corey, L. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection: A controlled trial in persons with fewer than 500 cd4-positive cells per cubic millimeter. *New England Journal of Medicine*, 322(14):941–949.