Scuola di Economia e Management
Corso di Laurea in Statistica, Scienze Attuariali e Finanziarie

Tesi di Laurea

Un Approccio di Screening Bayesiano Aggiustato per Covariate per Traiettorie di Biomarcatori Multipli

A Covariate-Adjusted Bayesian Screening Approach for Multiple Longitudinal Biomarkers

Maria Veronica Vinattieri

Relatore: *Francesco Claudio Stingo*
Correlatore: *Michela Baccini*

Anno Accademico 2018-2019

Contents

List of Figures

— *A Eleonora Lazzeri*

Abstract

Advanced hepatocellular carcinoma (HCC) has restricted treatment options and low survival, therefore early detection is crucial since even a little improvement may result a significant improvement in HCC patients survival rates. Recent developments in methods for early detection of HCC have focused on the analysis of trajectories of multiple biomarkers, assuming biomarkers to follow a joint hierarchical mixture model with random changepoints. We propose an innovative extension that consists in the inclusion of baseline covariates to the model on multiple longitudinal biomarkers trajectories. We want to assess whether covariates could capture a component of variation of biomarkers trajectories in order to improve early detection. A MCMC algorithm was conducted to derive posterior distributions and the posterior risk of being a case for each patient. The data from the Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) Trial contains valuable information about multiple longitudinal biomarkers AFP and DCP in cirrhosis patients with an extensive follow-up. The screening algorithm is applied in simulations studies under a range of possible scenarios and then in the HALT-C Trial using in-sample validation.

Chapter 1

Introduction

Advanced hepatocellular carcinoma (HCC) is the most common liver cancer that affects people worldwide. It has a poor survival primarily due to the lack of effective treatments for late stage patients, that represent the majority of cases at diagnosis. Early detection is critical for this kind of cancer, and even a little improvement may result a significant improvement in survival rates.

Recent developments in methods for early detection of HCC patients have focused on the analysis of trajectories of multiple biomarkers, assuming for them and in particular for their changepoints a joint model. Longitudinal biomarkers are indeed an essential tool for screening test, because they totally respect desirable characteristics of the test. Indeed, they are non-invasive in order to reduce patients anxiety and clinical costs and they are also inexpensive, to allow a widespread use.

The model proposed here is an innovative extension of the fully Bayesian hierarchical joint model (mFB) outlined in Tayob et al. [2018]. The extension consists of the inclusion of baseline covariates to the model on multiple biomarkers trajectories. The aim of the new proposed method is to find out whether pre-analysis clinical features in addition to biomarkers levels are able to improve early detection.

The accuracy of the new screening algorithm has to be evaluated before being applied in a prospective study. Therefore, simulation studies have been conducted to assess whether covariates could catch a component of variation of biomarkers trajectories, variation that is not explained in the newest approach in Tayob et al. [2018]. The hepatitis C antiviral long-term treatment against cirrhosis (HALT-C) trial contains valuable information

about multiple longitudinal biomarkers in cirrhosis patients with an extensive follow-up. The HALT-C dataset has indeed been used to evaluate the current screening algorithm once its accuracy has been ensured in the simulation phase.

This work will demonstrate that the new approach has a higher sensitivity than the old approach when covariates play an key role in predicting the biomarkers trajectories (section 5.1.3).

The thesis is organized as follows: in chapter 2 the background is described, in particular there is the description of HALT-C Trial data. In chapter 3 diagnostic tests are explained and their main features are described; in addition a first sketch of the new screening approach is given. In chapter 4 all the elements of the new screening test are outlined, including the joint model for covariates-adjusted multiple biomarkers and the computational procedure implemented to obtain necessary posterior distributions for the computation of the posterior risk of disease. Chapter 5 reports results from multiple scenarios simulation studies. Once the new screening approach has been assessed on simulation studies, it has been applied to real data. Results of applying the new screening approach on the HALT-C Trial are also reported in section 5. In chapter 6 a discussion is given.

Chapter 2

Background

This chapter starts with an introduction to HCC, a discussion on the relevance of early detection for this disease and on biomarkers potential. It is then described the HALT-C trial, a prospective study that has been implemented to collect and analyze two biomarkers of interest, AFP and DCP.

2.1 Hepatocellular Carcinoma

Hepatocellular carcinoma, the most common liver cancer, affects half a million people each year, with 20000 people in the United States alone. Liver cancer is in the top-10 most common cancers in both men and women, respectively at the fifth and seventh rank [El-Serag and Mason, 1999], [El-Serag, 2011].

Hepatocellular carcinoma incidence is higher in developing countries, where the transmission of the hepatitis B virus (HBV) is endemic, i.e. it is rooted in that particular regions from mother to newborns. Indeed, hepatocellular carcinoma is linked to infections by HBV or hepatitis C virus (HCV). The onset of those infections have become one the main rising causes of hepatocellular carcinoma.

Hepatocellular carcinoma occurs rarely before the age of 40, and it has its edge-risk at 70 years old. Moreover, the incidence raised in the past years among hispanics and white people. In the United States it has been seen that during the last 2 decades the incidence of hepatocellular carcinoma (HCC) increased 3 times, while the 5-year survival diseased patients were steadily less

than 12% [El-Serag and Mason, 1999].

It has been seen that HBV and HCV are two of the main risk factors to cause the occurrence of the hepatocellular carcinoma. The risk they bring depends on magnitude on the ethnic group and on geographic region. Overall, they cause the occurrence of cirrhosis. Indeed, the 80-90% of HCC cases have cirrhosis. In cirrhosis diseased patients the increasing risk of developing a hepatocellular carcinoma goes from 5% to 30% within 5 years. The risk variability depends on the cause of cirrhosis (HCV, HBV, fatty liver disease or others), on the geographic region and on the stage of cirrhosis.

Among hepatocellular carcinoma cases the 50% is infected by HBV. Moreover, the 70-80 % of HBV and hepacellular carcinoma diseased patients have also cirrhosis. Therefore it is likely that hepatocellular carcinoma, HBV and cirrhosis are strictly linked to each other. The risk of liver cancer onset in the HBV-infected patients varies on: gender, age, time of infection, family cancer history, use of alcohol or tobacco and the co-presence of HCV infection or hepatitis.

But the only presence of HBV is less risky than HCV. Indeed, HCV infection is the higher risk factor and it is at most dangerous with the co-presence of liver fibrosis. Risk of developing a hepatocellular carcinoma in HCV-infected patients is 15-20 times more likely than in HCV-free people. It has been seen that in Italy the 40-60% of liver cancer diseased are infected by HCV, while in the United States the percentage is up to the 50% and it is expected to increase in the next 3 decades. The variability of the risk of HCV-infection on cancer onset depends on age (the older the higher risk), gender (males have the highest risk), co-presence of HBV, diabetes or obesity, use of alcohol.

In the United States 30-40% of hepacellular carcinoma cases are not infected by neither HCV nor HBV. Therefore there are other risk causes. One of the main secondary causes could be the obesity factor. Hispanic and white races have been found to be likely to contract HCC the most. Lastly, drinking coffee has been luckily found to be related with a lower risk to develop liver cancer [El-Serag, 2011].

2.1.1 Early Detection and Biomarkers Potential

Early detection is an incredible powerful tool to improve survival of cancer patients. A patient diagnosed with an early stage cancer has multiple treatment

options and therefore the probability to recover from the disease is higher than late-stage detection patients. Especially, what it have been seen about HCC, early detection allows patients to find a successful treatment. Furthermore, early detected patients 5-years survival is over 60%, while later stage patients survival is up to 10% [Bruix and Sherman, 2005].

In order to successfully increase early detection, screening tests are used to identify the disease at the most curable stage. A screening test is a method of secondary prevention and it deals with the detection of a disease before the symptoms appearance. On the other hand, for a low-incidence disease it is frequent that a test results positive but it is truly not. That is why screening tests for early detection have the second aim of keeping low the false positive rate to avoid clinical and financial wastes [Skates et al., 2001].

The six-month ultrasonography is used in order to detect HCC but many problems arise using this method because ultrasonography is: operator dependent, not easily performed in obese patients, not sensitive for early lesions and there are significant problems about quality variability among hospital facilities [Tayob et al., 2018].

Biological markers - or biomarkers - offer an inexpensive non-invasive approach for screening, although the disadvantage is about their large intervals collection, that is annual. In respect to the disease onset they may behave differently: some biomarkers levels can rise exponentially, while others do not. Usually, continuous biomarkers are converted in dichotomous in order to establish a cut-off level to determine the disease status.

Usually the diagnostic HCC biomarker serum α-Fetoprotein (AFP) is integrated to ultrasonography. It has a notable sensitivity between 41% and 100%, and specificity between 70% and 95%. Moreover, Des-γ carboxy-prothrombin (DCP) has shown its potential as a complementary screening biomarker in detecting HCC [Tayob et al., 2018].

2.1.2 HALT-C trial

The target population for HCC surveillance is composed by cirrhosis patients. The Hepatitis C Antiviral Long-term Treatment (HALT-C) Trial aimed to prevent fibrosis progression in HCV-infection patients through the evaluation of an interferon-based therapy (Tayob et al. [2018]).

Patients underwent a long follow-up and were monitored to control the development of HCC. They were visited every 3 months for the first 42 months

and every 6 months thereafter. Patients underwent clinical tests, including AFP in local laboratories and DCP in a central laboratory, but just for the first 42 months. They occasionally had a liver ultrasound at 6, 18, 30, 42 months and every 6 months thereafter. Patients with new lesions on ultrasound or high level of AFP were further evaluated with computed tomography (CT) and magnetic resonance imaging (MRI).

HCC diagnosis was based on imaging with or without AFP in absence of histology. HCC has been evaluated on all the patients of the 2 treatment groups.

HCV-cirrhosis patients are at high risk to contract HCC therefore they are recommended for surveillance. The data consist in 48 HCC cases and 361 control patients without HCC. The median follow-up period is about 78 months.

Chapter 3

Statistical methods

Diagnostic tests are described in this chapter. Especially the concepts of sensitivity, specificity and ROC curve are outlined. Moreover, the computation of the predictive value of a test is given. Then, the new screening method is described.

3.1 Diagnostic tests

Diagnostic tests are able to detect patients recently affected by the disease of interest. A diagnostic test has some desirable characteristics that make it an ideal test. It is desirable to have it fast in the execution, safe and simple. Moreover, it is wanted to be as less invasive as possible, in the better cases painless. It is preferable to have it cheap and reliable.

A diagnostic test has a predictive variable that indicates the result of the test and a outcome variable that represents presence or absence of the disease.

The aim of diagnostic tests is to discriminate between cases and heath patients, given 4 different situations resulting form the combination of disease status (positive and negative) and test result (positive and negative). Everything is represented in the following table

	Diseased Patient (+)	Disease-free Patient(-)
Positive Test (+)	TP	FP
Negative Test (-)	FN	TN
Total	TP+FN	FP+TN

where TP means true positive, FP means falso positive, FN means false negative and TN means true negative.

Two measures are used to assess the goodness in prediction of a diagnostic test: sensitivity and specificity. Sensitivity represents the proportion of cases that truly have a positive test out of all the diseased patients

$$SE = \frac{TP}{TP + FN}$$

while specificity represents the proportion of healthy patients with a truly false test out of all the healthy patients

$$SP = \frac{TN}{TN + FP}.$$

When sensitivity is higher, specificity is lower. The contrary happens as well. It depends on the nature of disease (rare or common) which one of these 2 measures has to prevail: for a rare disease it is better to have less false positive so higher sensitivity. On the contrary, for a common disease false negative rate therefore higher specificity is better. Making mistakes about detecting a rare disease in truly disease-free patients could produce clinical and economic drawbacks. Therefore, being as much precise as possible is preferred. As well as identifying as much as possible cases with a not rare disease is desirable.

Test result can be continuous or discrete. When it is continuous a threshold has to be chosen to make decisions about the outcome. With the threshold choice it is possible to make sensitivity or specificity prevails (necessarily when one gets higher the other gets lower, and vice-versa).

The receiver operating characteristic (ROC) curve is another way to set the threshold to a certain value. The curve is given by all the combinations of sensitivity and specificity of the model by varying the threshold. It is possible to represent the ROC curve on a Cartesian plain with sensitivity on the y axis and 1 - specificity on the x axis. Sensitivity is the proportions of how many diseased have a positive test result, at the contrary (1 - specificity) is how many without the disease have a positive test result. The best result is when sensitivity is as close as possible to 1 and (1 - specificity) is as close as possible to 0: the ideal test results in the left-hand upper side of the graph. When it happens, the area under the ROC curve (AUC) is close to 1, that is the optimum point. A value of 0.5 is what it is expected to find randomly.

Each patient has a prior probability to have the disease before undertaking the test, with respect to demographic and clinical characteristics.

Given D^+ when the patient is diseased and D^- when the patient is healthy, T^+ and T^- when the test is respectively positive and negative, the predictive value of a positive test is defined as follows

$$PV^+ = P(D^+|T^+) = \frac{P(T^+|D^+) \cdot P(D^+)}{P(T^+|D^+) \cdot P(D^+) + P(T^+|D^-) \cdot P(D^-)}$$

where $P(T^+|D^+) = SE$ and $P(T^+|D^-) = 1 - SP$.

The predictive value of a positive test represents the probability of truly having the disease given positive result of the test itself .

While, the predictive value of a negative test is defined as

$$PV^- = P(D^-|T^-) = \frac{P(T^-|D^-) \cdot P(D^-)}{P(T^-|D^-) \cdot P(D^-) + P(T^-|D^+) \cdot P(D^+)}$$

where $P(T^-|D^-) = SP$ and $P(T^-|D^+) = 1 - SE$.

The predictive value of a negative test represents the probability of truly not having the disease given negative result of the test itself.

The most common method to carry out a diagnostic test consists in fixing a threshold and representing a ROC curve to assess the results. This method is simplistic and has some limitations:

1. a common threshold is chosen for all patients;
2. only one biomarker is usually used for the detection of a disease. No combination methods for multiple biomarkers is outlined;
3. one screening value is used to assess the disease status, that coincide with the last one measured. No screening values over time are used.

All these limitations allow to create a new approach of screening.

3.2 New screening test

A new approach of screening test is computed: a subject-specific threshold is chosen with respect to patients involved in the screening test. Multiple

biomarkers are used to detect the asymptomatic disease, and they are factors of a unique joint model. All patients are followed up over time and all the screening values are used to detect the onset disease, not only the last one in time.

One of the main aspect is how biomarkers are considered jointly because usually they behave in different ways when the disease onset shows up. The time point the biomarker trajectory changes is easily called the changepoint. In the method the similarity between changepoints of different biomarkers with respect to the same disease onset is assumed. Overall, an increase in biomarker level is considered indicative of a latent disease onset. On the contrary, a possible decrease in biomarker level can be easily adjusted. The behaviour of all measured biomarkers at the disease onset is considered jointly. The joint model of the changepoints allows to borrow information across the biomarkers in order to identify the changepoints more subtle. This happens when changepoints are not easily identifiable but the patient the changepoints belong to is diseased. Indeed, there are cases where only one biomarker trajectory is not enough to capture a changepoint: it means that some biomarkers are not enough to catch a fundamental possible signal of the disease onset (Chapter 5).

Chapter 4

Statistical Methods for Longitudinal Biomarkers

The method implemented in this work is a innovative contribution to the model in Tayob et al. [2018]. It is indeed a covariates-adjusted fully Bayesian hierarchical changepoints and mixture models of longitudinal biomarkers. The aim is to capture biomarkers trajectories, especially trajectories changepoints in order to detect HCC cases at the earlier stage, with the additional use of covariates.

In this chapter a literature review is made at the beginning, than the new screening approach is deeply outlined.

4.1 Literature review

Some of the developed screening methods about early detection of HCC are reported as the starting point of this work:

- McIntosh and Urban [2003], univariate parametric empirical Bayes
- Skates et al. [2001], fully Bayesian screening algorithm for a single longitudinal biomarker trajectory;
- Tayob et al. [2018], fully Bayesian screening algorithm for multiple longitudinal biomarkers.

All these methods are outlined in a very generalized way. The reason is that all methods work with any biomarker and can be applied to any asymptomatic disease. After outlining the already known approaches, the new method is explained in a more deep way and especially for this particular case - AFP, DCP on HCC detection.

4.1.1 Univariate parametric empirical Bayes

The idea McIntosh and Urban [2003] had was to implement a model that incorporated prior screening history of case patient and a model on biomarkers in control patients. The model is called univariate parametric empirical Bayes (uPEB).

Let Y_{ij1} be the biomarker level. The disease status is denoted by D and assumes one of 2 values for the $i-th$ patient: $D_i = 0$ when patient is disease-free at time d_i therefore the marker level varies randomly around its mean θ_{i1} following the model

$$Y_{ij1} \sim N(\theta_{i1}, \sigma_1^2)$$
$$\theta_{i1} \sim N(\mu_{\theta_1}, \sigma_{\theta 1})$$

where, given θ_{i1} the biomarker levels Y_{ij1} are independent and identical distributed. Within-subject variance σ_1^2 and between-subject variance $\sigma_{\theta 1}$ are key measures in this method. Biomarker levels can be standardized to simplifies the computation:

$$Z_{ij} = (Y_{ij1} - \mu_{\theta_1})/\sqrt{\sigma_1^2 + \sigma_{\theta 1}}$$
therefore
$$Z_{ij}|\mu_i \sim N(\mu_i, 1 - B_1)$$
$$\mu_i \sim N(0, B_1) \qquad \text{where } B_1 = \frac{\sigma_{\theta 1}}{\sigma_{\theta 1} + \sigma_1^2}$$

This method is an extension of the standard threshold (ST) screening approach that do not use the past screening history of each patient indeed a fixed threshold is chosen for all the patients.

The threshold is chosen in order to keep the false-positive rate (FPR) among control patients less than f_0. As Z_{ij} is distributed as a standard Normal, the

probability $P(Z_{ij} > z_{1-f_0}) = f_0$ where z_{1-f_0} is the 100(1-f_0) quantile of the standard Normal distribution. Therefore, a patient is considered diseased when its own Z_{ij} exceeds the threshold z_{1-f_0}

$$Z_{ij} > z_{1-f_0} \rightarrow \text{case patient.}$$

when μ_i is known for each patient, the threshold can be personalized. The probability is $(P(Z_{ij} - \mu_i)/\sqrt{1-B_1} > z_{1-f_0}|\mu_i) = f_0$, and a patient is considered diseased when its own Z_{ij} exceeds the threshold $\mu_i + z_{1-f_0}\sqrt{1-B_1}$

$$Z_{ij} > \mu_i + z_{1-f_0}\sqrt{1-B_1} \rightarrow \text{case patient}$$

Since μ_i is not known, it is estimated $\hat{\mu_{ij}}$ as a weighted mean between the population mean and the sample mean of the past screening history for each patient.

The rule to decide whether the patient is diseased is the following

$$Z_{ij} > \hat{\mu_{ij}} + z_{1-f_0}\sqrt{1-B_1B_j} \rightarrow \text{case patient}$$

where $\mu_{ij} = 0 \cdot (1 - B_j) + \bar{Z_{ij}} \cdot B_j$, with $\bar{Z_{ij}} = \dfrac{1}{j-1}\sum_{j'=1}^{j-1} Z_{ij'}$ and $B_j = \dfrac{\sigma_{\theta_1}}{\sigma_1^2/(j-1) + \sigma_{\theta_1}}$.

4.1.2 Univariate Bayesian

Skates et al. [2001] proposed a univariate fully Bayesian screening algorithm (uFB) for a unique longitudinal biomarker trajectory.

Model

Let Y_{ij1} be the marker level, where i indicates patient, j indicates screening time, t_{ij} indicates the visit time measured in years from the entry date, and 1 is for the unique biomarker index. The disease status is denoted by D that assumes the following 2 values for the $i-th$ patient:

- $D_i = 0$ when patient is disease-free at time d_i and the marker level varies

randomly around the mean θ_{i1} following the model

$$Y_{ij1} = \theta_{i1} + \epsilon_{ij1}$$

with Normal distributed errors.

- $D_i = 1$ when patient is diseased at time d_i. I_{i1} is an indicator for the existence of a changepoint and is defined to distinguish between the 2 cases when $D_i = 1$ occurs. If $I_{i1} = 0$ the marker level trajectory does not change after disease onset so it varies as randomly around the mean as in control patients. On the other side, if $I_{i1} = 1$ the marker level fluctuates randomly around a constant mean θ_{i1} until the disease onset at time τ_{i1}, whence the marker level is added of a linear rate γ_{i1} as the time increases from the changepoint, following the model

$$Y_{ij1} = \theta_{i1} + \gamma_{i1}(t_{ij} - \tau_{i1})^+ + \epsilon_{ij1} \tag{4.1}$$

with Normal distributed errors. $()^+$ is the positive part of the expression.

Therefore the assumption is that the marker level linearly increases after the disease onset; for this reason some appropriate transformations are made to adjust the decreasing trajectories after the disease onset.

Priors

- Variance $1/\sigma_1^2$ of biomarker has a uninformative Jeffreys' prior;
- biomarker level mean θ_{i1} has a Normal distribution $N(\mu_{\theta 1}, \sigma_{\theta 1}^2)$;
- random effect for rate γ_{i1} has a log-Normal distribution $log(\gamma_{i1}) \sim N(\mu_{\theta 1}, \sigma_{\theta 1}^2)$;
- changepoint time τ_{i1} has a truncated Normal distribution $N_{[d_i - \tau^*, d_i]}(d_i - \mu_{\tau 1}, \sigma_{\tau 1}^2)$; where τ^* is the fixed time when the disease begins to show up. This information comes from pre-clinical studies. For HCC it is assumed to be 2 years;
- binary indicators I_{i1} follows a Bernoulli distribution with parameter $\pi_{i1} = exp(\mu_I)/\{1 + exp(\mu_I)\}$. This is a reduced case of Markov

Random Field, the joint distribution of I_{ik} among multiple biomarkers (paragraphh 4.2.1).

Posterior Risk

The biomarker level has to be evaluated in order to decide whether it is linked to diseased patients or not. Given the posterior risk of disease defined as

$$\frac{P(D_i = 1|Y_{ij1})}{P(D_i = 0|Y_{ij1})} = \frac{P(Y_{ij1}|D_i = 1) \cdot P(D_i = 1)}{P(Y_{ij1}|D_i = 0) \cdot (1 - P(D_i = 1))}$$

where D_i is the disease status of the patient, and Y_{ij} the biomarker level.

The decision rule is based on the posterior risk of disease related to the current screening value compared to all the available screening values from previous studies.

4.1.3 Independent Fully Bayesian

The most recent model is an extension of Skates to multiple correlated longitudinal biomarkers. For a multiple biomarkers fully Bayesian screening test (mFB) a joint model is computed: single trajectories may or may not exhibit changes at the cancer onset and they have not the same changepoint time in function of the occurrences. However, in this method the similarity between changepoints with respect to the cancer onset is assumed.

Method

The method is the same as in Skates but with an additional index K in order to indicate different biomarkers.

Let Y_{ijk} be the biomarker level, where i indicates patient, j indicates screening time related to t_{ij} indicates visit time measured in years from the entry date, and k indicates the biomarker. Disease status is denoted by D and it assumes 2 values:

$D_i = 0$ when patient is disease-free at time d_i therefore the marker level varies randomly around mean θ_{ik} following the model

$$Y_{ijk} = \theta_{ik} + \epsilon_{ijk}$$

with Normal distributed errors;

while $D_i = 1$ when patient is diseased at time d_i. I_{ik} is defined to distinguish between 2 cases when $D_i = 1$. If $I_{ik} = 0$ the marker level trajectory does not change after disease onset so it varies as randomly as for control patients; while if $I_{ik} = 1$ the marker level fluctuates randomly around a constant mean θ_{ik} until the disease onset at time τ_{ik}, whence the marker level is added of a linear rate γ_{ik} as the time increases from the changepoint following the model

$$Y_{ijk} = \theta_{ik} + \gamma_{ik}(t_{ij} - \tau_{ik}) + \epsilon_{ijk}$$

.

Therefore the assumption is that the marker level linearly increases after the disease onset. Appropriate transformations are made to adjust the decreasing trajectories after the disease onset.

The algorithm is efficient because it can accommodate inconstant visits of patients that do not follow recommended surveillance and produce missing data.

This is a more efficient strategy of screening for early detection of low-incidence diseases than Skates strategy, still based on biomarkers. Multiple biomarkers screening is essential in order to produce high sensitivity test. The risk of HCC for future patients is not taken into account. The longitudinal trajectory of biomarkers is the main interest [Tayob et al., 2018].

Priors

Since the model is hierarchical then parameters are referred to 2 levels: patients and biomarkers. Patient level parameters allow to personalize the threshold for the risk computation.

Subject-specific parameters

- Biomarker level average θ_{ik} has a Normal distribution $N(\mu_{\theta k}, \sigma^2_{\theta k})$;
- random effect for rate γ_{ik} has a log-Normal distribution $log(\gamma_{ik}) \sim N(\mu_{\theta k}, \sigma^2_{\theta k})$;
- changepoint time τ_{ik} has a truncated Normal distribution $N_{[d_i-\tau^*,d_i]}(d_i - \mu_{\tau k}, \sigma^2_{\tau k})$; where d_i is the exit time and τ^* is the fixed

time when the disease begins to show up. This information comes from pre-clinical studies. For HCC it is fixed to 2 years.

- binary indicators $I_i = (I_{i1}, ..., I_{ik})$ follows a Markov Random Field (MFR) distribution

$$P(I_i) \propto exp\left\{\mu_I\left(\sum_{k=1}^{K} I_{ik}\right) + \eta_I(I_i^T R I_i)\right\}$$

where R is a upper triangular matrix that provides the correlation between changepoints, μ_I controls the sparsity of the model, η_I controls the smoothness of the ditribution of I_i, therefore it controls the changepoints dependency. Changepoints are assumed to be correlated to each other, then the probability of observing a changepoint in the k-th marker of the i-th patient depends on the discovered changepoints in the (k-1)-ths previous markers. The conditional distribution of I_i results

$$P(I_{ik}|(I_{ik'} : k' \neq k)) = \frac{exp\{I_{ik}F(I_{ik})\}}{1 + exp\{F(I_{ik})\}}$$
$$\text{where } F(I_{ik}) = \mu_I + \eta_I \sum_{k' \neq k} I_{ik'};$$

- $logit(\mu_I)$ has a Beta prior;
- η_I has a Beta prior.

Biomarker-specific parameters

- Variance $1/\sigma_k^2$ of each k biomarker has a uninformative Jeffreys' prior;
- Means $\mu_{\theta k},\ \mu_{\gamma k},\ \mu_{\tau k}$ have Normal priors;
- Variances $\sigma^2_{\theta k},\ \sigma^2_{\gamma k},\ \sigma^2_{\tau k}$ have Inverse Gamma priors.

Posterior Risk

The posterior joint model for all the parameters is not available in closed form, then a MCMC algorithm is computed to sample from the posterior

distributions. The full conditional of the biomarker-specific parameters are easily computable and a Gibbs sampler step is used, except for parameters related to the changepoint ($\mu_{\tau k},\ \sigma^2_{\tau k}$) and except for MFR parameters ($\mu_I,\ \eta_I$). Therefore, a Metropolis-Hastings is implemented for them.

A Gibbs sampler is used to sample from the full conditional (FC) of the subject-specific parameter θ_{ik} as well.

The posterior distributions of subject-specific parameters $I_{ik},\ \gamma_{ik},\ \tau_{ik}$ are connected and they are computed as follows: if $I_{ik} = 0$ there is only θ_{ik} distribution; if $I_{ik} = 1$ there are all parameters distributions. Therefore, the space of parameter depends on the value of I_{ik} and a reversible-jump step (section 4.2.3) is used to sample from the FC of all the subject-specific parameters.

The decision of the disease status of a (N+1)th patient at time t_{ij} is based on his posterior risk of disease, given the longitudinal trajectory of each biomarker until time t_{ij}.

The posterior risk is computed as follows

$$\frac{P(D_{N+1}=1|Y_{N+1})}{P(D_{N+1}=0|Y_{N+1})} = \frac{P(Y_{N+1}|D_{N+1}=1)}{P(Y_{N+1})|D_{N+1}=0} \cdot \frac{P(D_{N+1}=1)}{1-P(D_{N+1}=1)}$$
$$\text{where } Y_{N+1} = \{Y_{N+1j'k},\ j' = 1,...,j \text{ and } k = 1,...,K\}$$

Prior prevalence of disease $P(D_{N+1} = 1)$ can be estimated from previous surveillance on target population or from training data.

Conditional probabilities $P(Y_{N+1}|D_{N+1} = 1)$ and $P(Y_{N+1}|D_{N+1} = 0)$ are estimated through predictive distributions based on N patients biomarkers levels from training data. A Monte Carlo integration is used for this computation step.

If posterior risk exceeds a fixed threshold, patient has enough evidence to be a HCC case than to be a control. That means that the result of screening is positive and it is used with additional tests (CT, MRI) to ensure predicting the correct HCC disease status. Threshold depends on clinical context: in this case it is fixed to maintain low the false positive rate (FPR) in order to reduce costs and unnecessary anxiety.

Assessments

The accuracy of screening is given by sensitivity and specificity. Those concepts are extended to:

- *patient-level sensitivity* defined as the proportion of cases with at least one positive test during all the screening time;
- *screening-level specificity* defined as the proportion of negative tests out of all the tests undertaken on the control group.

4.2 Multiple Trajectories With Covariates

The model described in this section is an innovative extension of the fully Bayesian hierarchical joint model (mFB) outlined in Tayob et al. [2018]. The extension consists of the inclusion of baseline covariates to the model on multiple biomarkers trajectories. The model is hierarchical because of the 2 levels defined: the subject-specific level and the biomarker-specific level. Biomarkers changepoints are assumed to be correlated to each other. Biomarker level is assumed to vary randomly around a mean value until the disease onset, whence it may or may not change over time. It is assumed that an increase in biomarker level is signal of a latent disease.

Let Y_{ijk} be the $k-th$ biomarker level in the $i-th$ patient the $j-th$ screening time. Moreover, let X_{il} be the $l-th$ covariate for the $i-th$ patient, and β_{kl} be the regression coefficient associated with the $l-th$ covariate X_l for the $k-th$ biomarker. All other parameters are defined in the same way as in Tayob et al. [2018] (section 4.1.3) .

The disease status is denoted by D and can assume 2 values for the $i-th$ patient:

- $D_i = 0$ when the patient is disease-free at time d_i and the biomarker level varies randomly around average $\theta_{ik} + \beta_{kl}X_{il}$. The complete model is

$$Y_{ijk} = \theta_{ik} + \beta_{kl}X_{il} + \epsilon_{ijk}$$

whit Normal distributed errors.

- $D_i = 1$ when the patient is diseased at time d_i. I_{ik} indicates the existence of a changepoint and is defined to distinguish between 2 sub-cases. If $I_{ik} = 0$ the biomarker level trajectory does not change after disease onset therefore it fluctuates around its mean as randomly as in control patients. On the other side, if $I_{ik} = 1$ the marker level fluctuates randomly around a constant mean $\theta_{ik} + \beta_{kl}X_{il}$ until the disease onset at time τ_{ik}, whence the biomarker level is added of a linear rate γ_{ik} as the time increases from the changepoint, following the model

$$Y_{ijk} = \theta_{ik} + \beta_{kl}X_{il} + \gamma_{ik}(t_{ij} - \tau_{ik})^+ + \epsilon_{ijk}$$

  with Normal distributed errors. $()^+$ is the positive part of the expression.

It is made an assumption on the biomarker level linear increase after the disease onset. Appropriate transformations are made to adjust the decreasing trajectories after the disease onset [Tayob et al., 2018].

The separation between cases and controls allows to identify better changepoints and rate of change. The complete models are specified as

$$\begin{aligned} Y_{ijk}|t_{ij}, \{I_i = 1\} &\propto N(\theta_{ik} + \beta_{kl}X_{il} + \gamma_{ik}(t_{ij} - \tau_{ik})^+,\ \sigma_k^2) \\ Y_{ijk}|t_{ij}, \{I_i = 0\} &\propto N(\theta_{ik} + \beta_{kl}X_{il},\ \sigma_k^2) \end{aligned}$$

Without loss of generality, let $i = 1, \ldots, n_0$ be the index for controls in the study and let $i = n_0 + 1, \ldots, N$ be the index for case patients. The likelihood under the assumed model is

$$\begin{aligned} L(\mathbf{Y}; \mathbf{t}, \cdot) = & \prod_{i=1}^{n_0}\prod_{k=1}^{K}\prod_{j=1}^{J_i} \phi\left(\frac{Y_{ijk} - \theta_{ik} - \beta_{kl}X_{il}}{\sigma_k}\right) \\ & \times \prod_{i=n_0+1}^{N}\prod_{k=1}^{K}\prod_{j=1}^{J_i} \phi\left(\frac{Y_{ijk} - \theta_{ik} - \beta_{kl}X_{il}}{\sigma_k}\right)^{1-I_{ik}} \\ & \phi\left(\frac{Y_{ij'k} - \theta_{ik} - \beta_{kl}X_{il} - \gamma_{ik}(t_{ij} - \tau_{ik})^+}{\sigma_k}\right)^{I_{ik}}, \end{aligned}$$

where $\mathbf{Y} = \{Y_{ijk}, i = 1, \ldots, N,\ j = 1, \ldots, J_i$ and $k = 1, \ldots, K\}$, $\mathbf{t} = \{t_{ij}, i = 1, \ldots, N$ and $j = 1, \ldots, J_i\}$ and ϕ is the standard Normal probability density function [Tayob et al., 2018].

4.2.1 Priors

Model parameters are either subject-specific or biomarker-specific, because of the hierarchical nature of the model. Subject-specific level allows the risk computation to be personalized with respect to the patient clinical past history. The difference in sensitivity between this new approach and the fixed cut-off approach can be appreciate in Chapter 5. Priors are listed in the next paragraphs.

Subject-specific parameters

- Mean biomarker level θ_{ik} has a Normal distribution $N(\mu_{\theta k}, \sigma^2_{\theta k})$;
- random effect for rate γ_{ik} has a log-Normal distribution $log(\gamma_{ik}) \sim N(\mu_{\theta k}, \sigma^2_{\theta k})$;
- changepoint time τ_{ik} has a truncated Normal distribution $N_{[d_i-\tau^*,d_i]}(d_i-\mu_{\tau k}, \sigma^2_{\tau k})$; where τ^* is the fixed time when the disease begins to show up. This information comes from pre-clinical studies. For HCC is 2 years.
- binary indicators $I_i = (I_{i1}, ..., I_{ik})$ follows a Markov Random Field distribution (MFR)

$$P(I_i) \propto exp\left\{\mu_I\left(\sum_{k=1}^{K} I_{ik}\right) + \eta_I(I_i^T R I_i)\right\}$$

  All the elements of this distribution are outlined in Section 4.1.3. We just remind that changepoints are assumed to be correlated each other, then the probability of observing a changepoint in the $k-th$ marker of the $i-th$ patient depends on the discovered changepoints in the $(k-1)-th$ previous markers. The conditional distribution of I_i is computed:

$$P(I_{ik}|(I_{ik'} : k' \neq k)) = \frac{exp\left\{I_{ik}F(I_{ik})\right\}}{1 + exp\left\{F(I_{ik})\right\}}$$
$$\text{where } F(I_{ik}) = \mu_I + \eta_I \sum_{k' \neq k} I_{ik'};$$

- $logit(\mu_I)$ has a Beta prior;

- η_I has a Beta prior.

Biomarker-specific parameters

- Variance $1/\sigma_k^2$ of each $k-th$ biomarker has a uninformative Jeffreys' prior;
- means $\mu_{\theta k},\ \mu_{\gamma k},\ \mu_{\tau k}$ have Normal priors;
- variances $\sigma^2_{\theta k},\ \sigma^2_{\gamma k},\ \sigma^2_{\tau k}$ have Inverse Gamma priors;
- β_{kl} is Normally distributed $N(\beta_0,\ C\sigma^2_{\beta_k})$; where $\sigma^2_{\beta_k}$ is distributed as an Inverse Gamma $IG(a, b)$, where a and b are fixed values and C is a matrix $C = cI$, where I is the identity matrix.[1] It is assumed that baseline covariates are fixed over time and have the same effect for control and case patients.

It is difficult to understand the joint behaviour of all the subject-specific parameters, therefore I_i is the only parameter considered jointly among the biomarkers.

4.2.2 Markov Chain Monte Carlo

The posterior joint model for all parameters is not available in closed form, then a MCMC algorithm is computed to sample from the posterior distributions. The full conditional distributions (FC) of the biomarker-specific parameters are easily computable and a Gibbs sampler step is used, except for parameters related to the changepoint $(\mu_{\tau k},\ \sigma^2_{\tau k})$ and for MFR parameters $(\mu_I,\ \eta_I)$. The updating of these parameters is via a Metropolis-Hastings algorithm.

A Gibbs sampler is used to sample from the full conditional of the subject-specific parameter θ_{ik} as well. All Gibbs steps are fairly standards and are detailed in Section 4.4.2.

While, the posterior distributions of subject-specific parameters $I_{ik},\ \gamma_{ik},\ \tau_{ik}$ are connected and are computed as follows: if $I_{ik} = 0$ there is only θ_{ik} distribution; if $I_{ik} = 1$ there are all the parameters distributions. Therefore, the parameters space depends on the value of I_{ik} and a reversible-jump step is used to sample from the full conditionals of all the subject-specific parameters.

[1] We also tried the g-prior such as $C = cX'X$ but with poor results.

4.2.3 Reversible jump step

There are many cases where parameters space varies in his dimension, instead of fixing parameters at the beginning of the model specification. The reversible Markov chain samplers method [Skates et al., 2001] [Green, 1995] is an extension of the Metropolis-Hastings algorithm and allows to jump between parameters sub-spaces of different dimensions in a way that varying-dimension problems are solved.

Posteriors of random effect parameters (θ_i, γ_i) are straightforward because they are sampled from singular full conditionals. It is not the same for changepoint parameter τ_i, indeed a Metropolis-Hastings is used to approximate its posterior density. A reversible-jump is used in order to sample from the full conditional of I_i since its value implies different dimensions of the parameters space. The starting point is $I_i = 0$ for subject i. To propose a move to $I_i = 1$ the first thing is expanding the parameters space from θ_i to $(\theta_i, \gamma_i, \tau_i)$. The steps to compute are the following:

- a candidate γ^* is generated from its prior $log(\gamma_i) \propto N(\mu_\gamma, d_\gamma)$;
- a candidate τ^* is generated from its prior $\tau_i \propto N(d_i - a_\tau^2)I[d_i - b_\tau, d_i]$, where a_τ and b_τ are previously fixed;
- new parameters of the expanded space are $\theta^* = \theta_i$, γ^*, τ^*;
- the acceptance probability of the proposal move from $I_i = 0$ to $I_i = 1$ is: min {likelihood ratio $\times$ prior ratio $\times$ proposal ratio $\times$ Jacobian, 1}. Likelihood and prior ratios are evaluated at the values $(\theta_i, \gamma_i, \tau_i)$ under the model $I_i = 1$ over the value (θ_i) under the model $I_i = 0$; proposal ratio is the proposal density for moving from $I_i = 1$ to $I_i = 0$ over the proposal density for moving from $I_i = 0$ to $I_i = 1$ (in this case is 1 divided by the prior densities for γ_i and τ_i evaluated in γ_i^* and τ_i^*). It is also included the Jacobian of transformation from the current parameter space augmented by new parameters $(\theta_i, \gamma_i^*, \tau_i^*)$ to the new parameters space $(\theta_i^*, \gamma_i^*, \tau_i^*)$ (in this case the Jacobian=1 because $\theta_i = \theta_i^*$). Therefore, the current acceptance probability coincides with the Metropolis-Hastings acceptance probability.

4.2.4 Computational Algorithm

The algorithm is explained step by step in this section.

Algorithm

Step 0: Initialize parameters:

1. $\boldsymbol{\theta}_k^{(0)} = \{\theta_{ik}^{(0)},\, i = 1, \ldots, N\},\, k = 1, \ldots, K$
2. $\mu_{\theta k}^{(0)},\, k = 1, \ldots, K$
3. $\sigma_{\theta k}^{2(0)},\, k = 1, \ldots, K$
4. $\sigma_k^{2(0)},\, k = 1, \ldots, K$
5. $\mathbf{I}_k^{(0)} = \{I_{ik}^{(0)},\, i = n_0 + 1, \ldots, N\},\, k = 1, \ldots, K$
6. $\mu_I^{(0)}$
7. $\eta_I^{(0)}$
8. $\boldsymbol{\gamma}_k^{(0)} = \{\gamma_{ik}^{(0)}, i = n_0 + 1, \ldots, N : I_{ik} = 1\},\, k = 1, \ldots, K$
9. $\mu_{\gamma k},\, k = 1, \ldots, K$
10. $\sigma_{\gamma k}^2,\, k = 1, \ldots, K$
11. $\boldsymbol{\tau}_k^{(0)} = \{\tau_{ik}^{(0)}, i = n_0 + 1, \ldots, N : I_{ik} = 1\},\, k = 1, \ldots, K$
12. $\beta_{kl}^{(0)},\, k = 1, \ldots, K$ and $l = 1, \ldots, L$
13. $\sigma_{\beta_k}^{(0)},\, k = 1, \ldots, K$

Step 1-S: Update parameters for $s \in \{1, \ldots, S\}$ and $s^* = 1 + 3(s-1)$.

1. Update $\mu_{\theta k},\ k = 1, \ldots, K$: Sample $\mu_{\theta k}^{(s)}$ from $N(\mu_{0k^*}, \sigma_{0k^*}^2)$, where $\mu_{0k^*} = \dfrac{\sigma_{\theta k}^{2(s-1)}}{\sigma_{\theta k}^{2(s-1)} + N\sigma_{0k}^2}\mu_{0k} + \dfrac{\sigma_{0k}^2}{\sigma_{\theta k}^{2(s-1)} + N\sigma_{0k}^2}\sum_{i=1}^{N}\theta_{ik}^{(s-1)}$ and $\sigma_{0k^*}^2 = \dfrac{\sigma_{\theta k}^{2(s-1)}\sigma_{0k}^2}{\sigma_{\theta k}^{2(s-1)} + N\sigma_{0k}^2}$.

2. Update $\sigma^2_{\theta k}$, $k = 1, \ldots, K$: Sample $\sigma^{2(s)}_{\theta k}$ from $IG(a_{\theta^*_k}, b_{\theta^*_k})$, where $a_{\theta^*_k} = a_{\theta k} + N/2$ and $b_{\theta^*_k} = b_{\theta k} + \frac{1}{2}\sum_{i=1}^{N}(\theta^{(s-1)}_{ik} - \mu^{(s)}_{\theta k})^2$.

3. Update σ^2_k, $k = 1, \ldots, K$: Sample $\sigma^{2(s)}_k$ from $IG(a_{\sigma_k}, b_{\sigma_k})$, where $a_{\sigma k} = \frac{1}{2}\sum_{i=1}^{N} J_i$, $b_{\sigma k} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{J_i}(Y_{ijk} - \theta^*_{ijk} - \beta_{kl}X_{il})^2$ and $\theta^*_{ijk} =$
$$\begin{cases} \theta^{(s-1)}_{ik} & \text{if } D_i = 0 \text{ or } (D_i = 1 \text{ and } I^{(s^*-1)}_{ik} = 0) \\ \theta^{(s-1)}_{ik} + \gamma^{(s^*-1)}_{ik}(t_{ij} - \tau^{(s^*-1)}_{ik})^+ & \text{if } D_i = 1 \text{ and } I^{(s^*-1)}_{ik} = 1 \end{cases}$$

4. Update μ_I:

   (a) Generate μ^*_I from its proposal distribution $J_{\mu_I}(\mu_I|\mu^{(s-1)}_I) = N(\mu^{(s-1)}_I, \delta^2_{\mu_I})$.

   (b) Compute acceptance ratio

$$\begin{aligned} \log(r) &= \min\left[\log\left\{\frac{P(\mathbf{I}^{(s^*-1)}|\mu^*_I, \eta^{(s-1)}_I)P(\mu^*_I|p_1, p_2)}{P(\mathbf{I}^{(s^*-1)}|\mu^{(s-1)}_I, \eta^{(s-1)}_I)P(\mu^{(s-1)}_I|p_1, p_2)}\right\}, 0\right] \\ &= \min[\log\{P(\mathbf{I}^{(s^*-1)}|\mu^*_I, \eta^{(s-1)}_I)\} + \log\{P(\mu^*_I|p_1, p_2)\} \\ &\quad - \log\{P(\mathbf{I}^{(s^*-1)}|\mu^{(s-1)}_I, \eta^{(s-1)}_I)\} \\ &\quad - \log\{P(\mu^{(s-1)}_I|p_1, p_2)\}, 0] \end{aligned}$$

where $P(\mathbf{I}|\mu_I, \eta_I) = \prod_{i=n_0+1}^{n} \exp\left\{\mu_I \left(\sum_{k=1}^{K} I_{ik}\right) + \eta_I \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} I_{ik} I_{ik'}\right) \times c$

and $c^{-1} = \sum_{\tilde{I} \in I} \exp\left\{\mu_I \left(\sum_{k=1}^{K} \tilde{I}_k\right) + \eta_I \left(\sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \tilde{I}_k \tilde{I}_{k'}\right)\right\}$

for I = All possible combinations of vector $\tilde{I}$ with binary $\tilde{I}_k$

and $P(\mu_I|p_1, p_2) = P_{\text{Beta}}\left\{\frac{\exp(\mu_I)}{1+\exp(\mu_I)}\middle| p_1, p_2\right\} \left|\frac{d}{d\mu_I}\frac{\exp(\mu_I)}{1+\exp(\mu_I)}\right|$

$= P_{\text{Beta}}\left\{\frac{\exp(\mu_I)}{1+\exp(\mu_I)}\middle| p_1, p_2\right\} \frac{\exp(\mu_I)}{\{1+\exp(\mu_I)\}^2}$

(c) Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $\mu_I^{(s)} = \mu_I^*$; otherwise set $\mu_I^{(s)} = \mu_I^{(s-1)}$

5. Update η_I:

(a) Generate η_I^* from its proposal distribution $J_{\eta_I}(\eta_I|\eta_I^{(s-1)}) = \text{Beta}(\tilde{a}, \tilde{b})$, where $\tilde{a}$ and $\tilde{b}$ are chosen so that the mean and variance of the Beta distribution are $\eta_I^{(s-1)}$ and $\delta_{\eta_I}^2$ respectively.

(b) Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\mathbf{I}^{(s^*-1)}|\mu_I^{(s)}, \eta_I^*)P(\eta_I^*|p_3, p_4)}{P(\mathbf{I}^{(s^*-1)}|\mu_I^{(s)}, \eta_I^{(s-1)})P(\eta_I^{(s-1)}|p_3, p_4)} \cdot \frac{J_{\eta_I}(\eta_I^{(s-1)}|\eta_I^*)}{J_{\eta_I}(\eta_I^*|\eta_I^{(s-1)})}\right\}, 0\right]$$

$$= \min[\log\{P(\mathbf{I}^{(s^*-1)}|\mu_I^{(s)}, \eta_I^*)\} + \log\{P(\eta_I^*|p_3, p_4)\} - \log\{P(\mathbf{I}^{(s^*-1)}|\mu_I^{(s)}, \eta_I^{(s-1)})\} - \log\{P(\eta_I^{(s-1)}|p_3, p_4)\} + \log\{J_{\eta_I}(\eta_I^{(s-1)}|\eta_I^*)\} - \log\{J_{\eta_I}(\eta_I^*|\eta_I^{(s-1)})\}, 0]$$

(c) Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $\eta_I^{(s)} = \eta_I^*$;

otherwise set $\eta_I^{(s)} = \eta_I^{(s-1)}$

6. Update $\mu_{\gamma k}$, $k = 1, \ldots, K$: Sample $\mu_{\gamma k}^{(s)}$ from $N(\mu_{1k^*}, \sigma_{1k^*}^2)$, where $\mu_{1k^*} = \frac{\sigma_{\gamma k}^{2(s-1)}}{\sigma_{\gamma k}^{2(s-1)} + n_{Ik}\sigma_{1k}^2}\mu_{1k} + \frac{\sigma_{1k}^2}{\sigma_{\gamma k}^{2(s-1)} + n_{Ik}\sigma_{1k}^2}\sum_{i=1}^{n_{Ik}} \log(\gamma_{ik}^{(s^*-1)})$, $\sigma_{1k^*}^2 = \frac{\sigma_{\gamma k}^{2(s-1)}\sigma_{1k}^2}{\sigma_{\gamma k}^{2(s-1)} + n_{Ik}\sigma_{1k}^2}$ and $n_{Ik} = \sum_{i=n_0+1}^{N} I_{ik}^{(s^*-1)}$.

7. Update $\sigma_{\gamma k}^2$, $k = 1, \ldots, K$: Sample $\sigma_{\gamma k}^{2(s)}$ from $IG(a_{\gamma_k^*}, b_{\gamma_k^*})$, where $a_{\gamma_k^*} = a_{\gamma k} + n_{Ik}/2$, $b_{\gamma_k^*} = b_{\gamma k} + \frac{1}{2}\sum_{i=1}^{n_{Ik}}\{\log(\gamma_{ik}^{(s^*-1)}) - \mu_{\gamma k}^{(s)}\}^2$ and $n_{Ik} = \sum_{i=n_0+1}^{N} I_{ik}^{(s^*-1)}$.

8. Update $\mu_{\tau k}$, $k = 1, \ldots, K$:

   (a) Generate $\mu_{\tau k}^*$ from its proposal distribution $J_{\mu_{\tau k}}(\mu_{\tau k}|\mu_{\tau k}^{(s-1)}) = N(\mu_{\tau k}^{(s-1)}, \delta_{\mu_{\tau k}}^2)$.

   (b) Compute acceptance ratio

   $$\log(r) = \min\left[\log\left\{\frac{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^*, \sigma_{\tau k}^{2(s-1)})P(\mu_{\tau k}^*|\mu_{2k}, \sigma_{2k}^2)}{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s-1)}, \sigma_{\tau k}^{2(s-1)})P(\mu_{\tau k}^{(s-1)}|\mu_{2k}, \sigma_{2k}^2)}\right\}, 0\right]$$
   $$= \min[\log\{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^*, \sigma_{\tau k}^{2(s-1)})\} + \log\{P(\mu_{\tau k}^*|\mu_{2k}, \sigma_{2k}^2)\}$$
   $$- \log\{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s-1)}, \sigma_{\tau k}^{2(s-1)})\} - \log\{P(\mu_{\tau k}^{(s-1)}|\mu_{2k}, \sigma_{2k}^2)\}, 0]$$

   (c) Generate $u \sim \text{Uniform}(0, 1)$. If $\log(u) < \log(r)$ set $\mu_{\tau k}^{(s)} = \mu_{\tau k}^*$; otherwise set $\mu_{\tau k}^{(s)} = \mu_{\tau k}^{(s-1)}$

9. Update $\sigma_{\tau k}^2$, $k = 1, \ldots, K$:

   (a) Generate $\sigma_{\tau k}^{2*}$ from its proposal distribution $J_{\sigma_{\tau k}^2}(\sigma_{\tau k}^2|\sigma_{\tau k}^{2(s-1)}) = TN_{[0,\infty]}(\sigma_{\tau k}^{2(s-1)}, \delta_{\sigma_{\tau k}}^2)$.

(b) Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s)}, \sigma_{\tau k}^{2*})P(\sigma_{\tau k}^{2*}|a_{\tau k}, b_{\tau k})}{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s)}, \sigma_{\tau k}^{2(s-1)})P(\sigma_{\tau k}^{2(s-1)}|a_{\tau k}, b_{\tau k})} \cdot \frac{J_{\sigma_{\tau k}^2}(\sigma_{\tau k}^{2(s-1)}|\sigma_{\tau k}^{2*})}{J_{\sigma_{\tau k}^2}(\sigma_{\tau k}^{2*}|\sigma_{\tau k}^{2(s-1)})}\right\}, 0\right]$$
$$= \min[\log\{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s)}, \sigma_{\tau k}^{2*})\} + \log\{P(\sigma_{\tau k}^{2*}|a_{\tau k}, b_{\tau k})\}$$
$$- \log\{P(\boldsymbol{\tau}_k^{(s^*-1)}|\mu_{\tau k}^{(s)}, \sigma_{\tau k}^{2(s-1)})\} - \log\{P(\sigma_{\tau k}^{2(s-1)}|a_{\tau k}, b_{\tau k})\}$$
$$+ \log\{J_{\sigma_{\tau k}^2}(\sigma_{\tau k}^{2(s-1)}|\sigma_{\tau k}^{2*})\} - \log\{J_{\sigma_{\tau k}^2}(\sigma_{\tau k}^{2*}|\sigma_{\tau k}^{2(s-1)})\}, 0]$$

(c) Generate $u \sim \text{Uniform}(0, 1)$. If $\log(u) < \log(r)$ set $\sigma_{\tau k}^{2(s)} = \sigma_{\tau k}^{2*}$; otherwise set $\sigma_{\tau k}^{2(s)} = \sigma_{\tau k}^{2(s-1)}$

10. Update each θ_{ik}, $i = 1, \ldots, N$ and $k = 1, \ldots, K$:

- If $D_i = 0$, sample $\theta_{ik}^{(s)}$ from $N(\mu_{\theta k^*}, \sigma_{\theta k^*}^2)$, where $\mu_{\theta_k^*} = \frac{\sigma_k^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}}\mu_{\theta k}^{(s)} + \frac{\sigma_{\theta k}^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}} \times \sum_{j=1}^{J_i}(Y_{ijk} - \beta_{kl}X_{il})$ and $\sigma_{\theta_k^*} = \frac{\sigma_k^{2(s)}\sigma_{\theta k}^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}}$.
- If $D_i = 1$ and $I_{ik}^{(s^*-1)} = 0$, sample $\theta_{ik}^{(s)}$ from $N(\mu_{\theta k^*}, \sigma_{\theta k^*}^2)$, where $\mu_{\theta_k^*} = \frac{\sigma_k^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}}\mu_{\theta k}^{(s)} + \frac{\sigma_{\theta k}^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}} \sum_{j=1}^{J_i}(Y_{ijk} - \beta_{kl}X_{il})$ and $\sigma_{\theta_k^*}^2 = \frac{\sigma_k^{2(s)}\sigma_{\theta k}^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}}$.
- If $D_i = 1$ and $I_{ik}^{(s^*-1)} = 1$, sample $\theta_{ik}^{(s)}$ from $N(\mu_{\theta k^*}, \sigma_{\theta k^*}^2)$, where $\mu_{\theta_k^*} = \frac{\sigma_k^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}}\mu_{\theta k}^{(s)} + \frac{\sigma_{\theta k}^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_{\theta k}^{2(s)}} \sum_{j=1}^{J_i}\{Y_{ijk} - \beta_{kl}X_{il} - \gamma_{ik}^{(s^*-1)}(t_{ij} - \tau_{ik}^{(s^*-1)})^+\}$ and $\sigma_{\theta_k^*}^2 = \frac{\sigma_k^{2(s)}\sigma_\theta^{2(s)}}{\sigma_k^{2(s)} + J_i\sigma_\theta^{2(s)}}$.

11. Update $\mathbf{I}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$.

(a) Update each I_{ik}, $i = n_0 + 1, \ldots, N$ and $k = 1, \ldots, K$:
If $I_{ik}^{(s^*-1)} = 0$,

i. Generate γ_{ik}^* from its prior $\log(\gamma_{ik}) \sim N(\mu_{\gamma k}^{(s)}, \sigma_{\gamma k}^{2(s)})$

ii. Generate τ_{ik}^* from its prior $\tau_{ik} \sim TN_{[d_i - \tau_k^*, d_i]}(d_i - \mu_{\tau k}^{(s)}, \sigma_{\tau k}^{2(s)})$

iii. Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\mathbf{Y}_{ik}|I_{ik}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^*,\tau_{ik}^*,\beta_{kl}^{(s)})}{P(\mathbf{Y}_{ik}|I_{ik}=0,\theta_{ik}^{(s)},\sigma_k^{2(s)}),\beta_{kl}^{(s)}} \cdot \frac{\pi_{ik}}{1-\pi_{ik}}\right\}, 0\right]$$

$$= \min[\log\{P(\mathbf{Y}_{ik}|I_i=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^*,\tau_{ik}^*),\beta_{kl}^{(s)}\} - \log\{P(\mathbf{Y}_{ik}|I_{ik}=0,\theta_{ik}^{(s)},\sigma_k^{2(s)}),\beta_{kl}^{(s)}\} + \log\{\pi_{ik}\} - \log\{1-\pi_{ik}\}, 0]$$

$$\text{where}\pi_{ik} = \frac{\exp\left\{\mu_I^{(s)} + \eta_I^{(s)}\left(\sum_{k'<k} I_{ik'}^{(s^*)} + \sum_{k'>k} I_{ik'}^{(s^*-1)}\right)\right\}}{1+\exp\left\{\mu_I^{(s)} + \eta_I^{(s)}\left(\sum_{k'<k} I_{ik'}^{(s^*)} + \sum_{k'>k} I_{ik'}^{(s^*-1)}\right)\right\}}$$

iv. Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $I_{ik}^{(s^*)} = 1$, $\gamma_{ik}^{(s^*)} = \gamma_{ik}^*$, and $\tau_{ik}^{(s^*)} = \tau_{ik}^*$; otherwise set $I_{ik}^{(s^*)} = 0$.

If $I_{ik}^{(s^*-1)} = 1$,

i. Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\mathbf{Y}_{ik}|I_{ik}=0,\theta_{ik}^{(s)},\sigma_k^{2(s)},\beta_{kl}^{(s)})}{P(\mathbf{Y}_{ik}|I_{ik}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*-1)},\tau_{ik}^{(s^*-1)},\beta_{kl}^{(s)})} \cdot \frac{1-\pi_{ik}}{\pi_{ik}}\right\}, 0\right]$$

$$= \min[\log\{P(\mathbf{Y}_{ik}|I_{ik}=0,\theta_{ik}^{(s)},\sigma_k^{2(s)},\beta_{kl}^{(s)})\} - \log\{P(\mathbf{Y}_{ik}|I_{ik}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*-1)},\tau_{ik}^{(s^*-1)},\beta_{kl}^{(s)})\} + \log\{1-\pi_{ik}\} - \log\{\pi_{ik}\}, 0]$$

$$\text{where } \pi_{ik} = \frac{\exp\left\{\mu_I^{(s)} + \eta_I^{(s)}\left(\sum_{k'<k} I_{ik'}^{(s^*)} + \sum_{k'>k} I_{ik'}^{(s^*-1)}\right)\right\}}{1+\exp\left\{\mu_I^{(s)} + \eta_I^{(s)}\left(\sum_{k'<k} I_{ik'}^{(s^*)} + \sum_{k'>k} I_{ik'}^{(s^*-1)}\right)\right\}}$$

ii. Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $I_{ik}^{(s^*)} = 0$, otherwise set $I_{ik}^{(s^*)} = 1$, $\gamma_{ik}^{(s^*)} = \gamma_{ik}^{(s^*-1)}$, and $\tau_{ik}^{(s^*)} = \tau_{ik}^{(s^*-1)}$.

(b) Update each γ_{ik}, $i \in \{i = n_0 + 1, \ldots, N : I_{ik}^{(s^*)} = 1\}$ and $k = 1, \ldots, K$:

i. Generate $\log(\gamma_{ik}^*)$ from its proposal distribution $J_{\gamma_{ik}}(\gamma_{ik}|\gamma_{ik}^{(s^*)}) = N(\log(\gamma_{ik}^{(s^*)}), \delta_{\gamma k}^2)$.

ii. Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^*,\tau_{ik}^{(s^*)},\beta_{kl}^{(s)})}{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*)},\tau_{ik}^{(s^*)},\beta_{kl}^{(s)})} \cdot \frac{P(\gamma_{ik}^*|\mu_{\gamma k}^{(s)},\sigma_{\gamma k}^{2(s)})}{P(\gamma_{ik}^{(s^*)}|\mu_{\gamma k}^{(s)},\sigma_{\gamma k}^{2(s)})} \frac{J_{\gamma_{ik}}(\gamma_{ik}^{(s^*)}|\gamma_{ik}^*)}{J_{\gamma_{ik}}(\gamma_{ik}^*|\gamma_{ik}^{(s^*)})}\right\}, 0\right]$$

$$= \min[\log\{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^*,\tau_{ik}^{(s^*)},\beta_{kl}^{(s)})\} - \log\{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*)},\tau_{ik}^{(s^*)},\beta_{kl}^{(s)})\} + \log\{P(\gamma_{ik}^*|\mu_{\gamma k}^{(s)},\sigma_{\gamma k}^{2(s)})\} - \log\{P(\gamma_{ik}^{(s^*)}|\mu_{\gamma k}^{(s)},\sigma_{\gamma k}^{2(s)})\} + \log\{J_{\gamma_{ik}}(\gamma_{ik}^{(s^*)}|\gamma_{ik}^*)\} - \log\{J_{\gamma_{ik}}(\gamma_{ik}^*|\gamma_{ik}^{(s^*)})\}, 0]$$

iii. Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $\gamma_{ik}^{(s^*+1)} = \gamma_{ik}^*$, $I_{ik}^{(s^*+1)} = 1$ and $\tau_{ik}^{(s^*+1)} = \tau_{ik}^{(s^*+1)}$; otherwise set $\gamma_{ik}^{(s^*+1)} = \gamma_{ik}^{(s^*)}$, $I_{ik}^{(s^*+1)} = 1$ and $\tau_{ik}^{(s^*+1)} = \tau_{ik}^{(s^*)}$.

(c) Update each τ_{ik}, $i \in \{i = n_0 + 1, \ldots, N : I_{ik}^{(s^*+1)} = 1\}$ and $k = 1, \ldots, K$

i. Generate τ_{ik}^* from its proposal distribution $J_{\tau_{ik}}(\tau_{ik}|\tau_{ik}^{(s^*+1)}) = TN_{[d_i-\tau_k^*,d_i]}(\tau_{ik}^{(s^*+1)}, \delta_{\tau k}^2)$.

ii. Compute acceptance ratio

$$\log(r) = \min\left[\log\left\{\frac{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*+1)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*+1)},\tau_{ik}^*,\beta_{kl}^{(s)})}{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*+1)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*+1)},\tau_{ik}^{(s^*+1)},\beta_{kl}^{(s)})}\cdot\right.\right.$$
$$\left.\left.\frac{P(\tau_{ik}^*|\mu_{\tau k}^{(s)},\sigma_{\tau k}^{2(s)})}{P(\tau_{ik}^{(s^*+1)}|\mu_{\tau k}^{(s)},\sigma_{\tau k}^{2(s)})}\frac{J_{\tau_{ik}}(\tau_{ik}^{(s^*+1)}|\tau_{ik}^*)}{J_{\tau_{ik}}(\tau_{ik}^*|\tau_{ik}^{(s^*+1)})}\right\},0\right]$$
$$= \min[\log\{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*+1)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*+1)},\tau_{ik}^*,\beta_{kl}^{(s)})\}$$
$$-\log\{P(\mathbf{Y}_{ik}|I_{ik}^{(s^*+1)}=1,\theta_{ik}^{(s)},\sigma_k^{2(s)},\gamma_{ik}^{(s^*+1)},\tau_{ik}^{(s^*+1)},\beta_{kl}^{(s)})\}$$
$$+\log\{P(\tau_{ik}^*|\mu_{\tau k}^{(s)},\sigma_{\tau k}^{2(s)})\}-\log\{P(\tau_{ik}^{(s^*+1)}|\mu_{\tau k}^{(s)},\sigma_{\tau k}^{2(s)})\}$$
$$+\log\{J_{\tau_{ik}}(\tau_{ik}^{(s^*+1)}|\tau_{ik}^*)\}-\log\{J_{\tau_{ik}}(\tau_{ik}^*|\tau_{ik}^{(s^*+1)})\},0]$$

iii. Generate $u \sim \text{Uniform}(0,1)$. If $\log(u) < \log(r)$ set $\tau_{ik}^{(s^*+2)} = \tau_{ik}^*$, $\gamma_{ik}^{(s^*+2)} = \gamma_{ik}^{(s^*+1)}$ and $I_{ik}^{(s^*+2)} = 1$; otherwise set $\tau_{ik}^{(s^*+2)} = \tau_{ik}^{(s^*+1)}$, $\gamma_{ik}^{(s^*+2)} = \gamma_{ik}^{(s^*+1)}$ and $I_{ik}^{(s^*+2)} = 1$.

12. Update each β_{kl}, $k = 1,\ldots,K$ number of biomarkers, $l = 1,\ldots,L$ number of covariates. Sample β_k from $MN(\mu_{\beta_k},\Sigma_{\beta_k})$, where

$$\mu_{\beta_k} = (C^{-1}/\sigma_{\beta_k}^{2(s-1)} + X^TX/\sigma_k^{2(s)})^{-1}X^Ty^*/\sigma_k^{2(s)}$$
$$\Sigma_{\beta_k} = (C^{-1}/\sigma_{\beta_k}^{2(s-1)} + X^TX/\sigma_k^{2(s)})^{-1}$$

with $y^* = Y_{ijk} - \theta_{ik}^{(s)} - \gamma_{ik}^{(s^*)}(t_{ij} - \tau_{ik}^{(s^*)})^+$ if $D_i = 1$ and $I_{ik}^{(s^*)} = 1$, and $y^* = Y_{ijk} - \theta_{ik}^{(s)}$ if $D_i = 0$ or $D_i = 1$ and $I_{ik}^{(s^*)} = 0$.

13. Update each σ_{β_k}, $k = 1,\ldots,K$. The full conditional is a Inverse-Gamma distribution with parameters $a + l^*/2$ - where l^* is the total number of covariates - and $b + (\beta_k^{(s)})^TC^{-1}\beta_k^{(s)}/2$.

4.2.5 Posterior Risk of Disease

The decision of the disease status of a (N+1)th patient at time t_{ij} is based on his posterior risk of disease, given the longitudinal trajectory of each biomarker until time t_{ij}.

The posterior risk is computed as

$$\frac{P(D_{N+1}=1|Y_{N+1})}{P(D_{N+1}=0|Y_{N+1})} = \frac{P(Y_{N+1}|D_{N+1}=1)}{P(Y_{N+1}|D_{N+1}=0)} \cdot \frac{P(D_{N+1}=1)}{1-P(D_{N+1}=1)}$$
$$\text{where } Y_{N+1} = \{Y_{N+1j'k},\ j' = 1, ..., j \text{ and } k = 1, ..., K\}$$

Prior prevalence of disease $P(D_{N+1} = 1)$ can be estimated from previous surveillance on the target population, or from training data as it has been done in the current work and in Tayob et al. [2018].

Conditional probabilities $P(Y_{N+1}|D_{N+1} = 1)$ and $P(Y_{N+1}|D_{N+1} = 0)$ are estimated through predictive distributions based on N patients biomarkers levels from training data.

A threshold is fixed to decide whether or not the patient is likely to be a case. First, a 90% level of specificity is fixed (i.e. the 90% of negative tests belong truly to disease-free patients), therefore the cut-off coincides with the 90% quantile - or for other values (specificity $\times$ 100)% quantile - of the posterior risk distribution computed on disease-free patients from training set. If posterior risk exceeds the fixed threshold, patient has enough evidence to be a HCC case than to be a control. That means that the screening result is positive and it is used with additional tests (CT, MRI) to ensure detecting the correct HCC disease status. The threshold depends on clinical context: in this case it is fixed in order to maintain low the false positive rate (FPR) in order to reduce costs, complications and unnecessary anxiety [Tayob et al., 2018].

4.2.6 Assessing accuracy

Accuracy of screening is given by sensitivity and specificity and by a graphical representation through the ROC curve. The concepts are extended to:

- *patient-level sensitivity* defined as the proportion of cases with at least one positive test during all the screening time;
- *screening-level specificity* defined as the proportion of negative tests out of all the tests undertaken on the control group Tayob et al. [2018].

Chapter 5

Results

This chapter shows the results from simulation studies and real data studies. The aim of simulation studies is to explore the performance of the new screening method under different scenarios of multiple biomarkers trajectories. Once the new screening has been assessed as accurate, it is applied to HALT-C Trial data.

5.1 Simulation studies

The aim of simulation studies is to compare the different screening methods to evaluate their potential in improving early HCC detection. The comparison is made on the evaluation of patient-level sensitivity corresponding to screening-level specificity, in the receiver operating characteristic (ROC) curve.

Training and validation data have been simulated to represent the aspects of the HALT-C trial data structure. Each group has 400 patients followed longitudinally for up to 5 years. The screening visits at time t_{ij} have been undertaken every 6 months, with variability due to the patients behaviour. Indeed, the number of screening visits J_i differs among patients.

Biomarkers levels are simulated from the joint model of controls and cases, with and without changepoint.

The covariate-adjusted model is then compared to the covariate-free model, that coincides with the mFB approach [Tayob et al., 2018]. In the scenario with non-informative covariates, the covariate-adjusted screening has a lower ROC than mFB screening (section 5.1.3). On the contrary, the new method improves

in detection than mFB when covariates assume much more importance.

5.1.1 Simulated data

The data that have been used in this section are an accurate simulation of the HALT-C trial data. The number of patients is fixed to 400. The disease status D is sampled from a Binomial distribution with probability 50/400 to be a case, that is a similar proportion of the one from HALT-C trial data. The time of the follow-up "d" is set maximum to 5 years and it is sampled for a Uniform[0,5] distribution. The time of visits are simulated on the following scheme: biomarkers are measured every 6 months, so the maximum number of visits that a patient can undertake is 11 in 5 years. The 4400 measurement times are sampled from a Normal distribution $N \sim (0.5, 0.1^2)$ and positioned in a matrix 400 patients $\times$ 11 measurements time. It is assumed the hypothesis of the human irregularity on having the visits, hence the variability of the visits is introduced in the data generation. In the first visit the biomarker level is fixed at 0. Each new visit has a time that indicates how much has passed since the first visit, therefore it is computed as a cumulative sum of the past times until that visit. The times of the visits after the exit time "d" are omitted because there are not post-diagnosis biomarker measurements. For each patient we know how many times they have been visited and the time intervals between the measurements from the very first time (time 0) till the exit time (d). For each patient a data-set is initialized in the following way:

1. "ID", status of disease "D" and exit time "d" are repeated for the number of visits undertaken;

2. cumulative time of the visits "t" are added to the dataset as a column;

3. total number of visits "J" are repeated for the number of visits undertaken;

4. a vector that counts the visits time after time "obs_number" is generated.

An example is shown in the table below.

ID	D	d	t	J	obs_number
351	0	2.483426	0.0000000	6	1
351	0	2.483426	0.3125795	6	2
351	0	2.483426	0.7785136	6	3
351	0	2.483426	1.2146442	6	4
351	0	2.483426	1.8363486	6	5
351	0	2.483426	2.3773203	6	6

Values for μ_I and η_I are set to a similar value from the HALT-C trial data-set and changepoints for the 3 biomarkers "I" are sampled, only for cases, from a Binomial distribution with a probability of being 1 that is the presence of a changepoint. Since the model is joint, the probability of changepoint changes.

Given $c_{inv} = 1 + 3exp(\mu_I) + 3exp(2\mu_I + \eta_I) + exp(3\mu_I + 3\eta_I)$

- $P(I_1 = 1) = \dfrac{(exp(\mu_I) + 2exp(2\mu_I + \eta_I) + exp(3\mu_I + 3\eta_I))}{c_{inv}}$
- $P(I_2 = 1|I_1) = \dfrac{\dfrac{(exp(\mu_I * (I_1 + 1) + \eta_I * I_1) + exp(\mu_I * (I_1 + 2) + eta_I * (2 * I_1 + 1)))}{c_{inv}}}{P(I_1 = 1)I_1 + (1 - P(I_1 = 1))(1 - I_1)}$
- $P(I_3 = 1|I_1, I_2) = \dfrac{exp(\mu_I + \eta_I * (I_1 + I_2))}{1 + exp(\mu_I + \eta_I * (I_1 + I_2))}$

Values for parameters used to simulate the data are set to values in the way to the best of reproducing Halt-C trial data-set. Values for σ_k^2, $\delta_{\sigma_k^2}$, μ_{θ_k}, σ_{θ_k}, μ_{γ_k}, σ_{γ_k}, $\mu_{\tau_k}, \sigma_{\tau_k}$ are taken from the paper of Tayob et al. [2018]. Values of the intercept θ_k are sample from a Normal distribution. Values of the linear rate γ_k are sampled, only for cases, from a Normal distribution and then they are raised to the exponential. Values of the changepoint time τ_k are sampled, only for cases, from a truncated Normal distribution.

Covariates are set to 2 and are randomly sampled from a Normal standard distribution $N(0, 1^2)$. All β parameters are equal with respect to the biomarkers and assume values in the interval [0, 2], depending on the simulation.

Mean and variance of biomarker level are computed via the formulas: $\bar{Y_k} = \theta_k + \beta X + \gamma_k(t - \tau_k)^{(+)}$, $sd(Y_k) = \sigma_k^2 + \delta_{\sigma_k^2}(t - \tau_k)^{(+)}$. Biomarker level coincides with its mean added by an uncertainty factor $Y_k = \bar{Y_k} + N(0, sd(Y_k))$.

Two times data are generated to obtain training dataset and validation dataset to assess the accuracy of the screening test.

5.1.2 Hyperparameter setting and sensitivity

In this section sensitivity of posterior inference on the novel parameters β is explored in comparison to results in Tayob et al. [2018]. All the other parameters are set to the values suggested by Tayob et al. [2018].

In the covariate-adjusted method $C = c(X'X)^{-1}$ or $C = cI$. Therefore, the value of c can be chosen to derive the best results in terms of posterior distribution of β and sensitivity. Four different cases have been computed in the way of choosing the best one.

Results with $C = c(X'X)^{-1},\ c = 0.1$ or $C = c(X'X)^{-1},\ c = 0.01$ are the worst because β covariance matrix has too small values, therefore this case can not represent well the true reality of the data. So c=1 is chosen. One of the 2 alternatives $C = I$ or $C = (X'X)^{-1}$ has to be chosen. They are therfore compared.

Scenario with $\beta = 2$ is used for generating data and comparing the 2 values of C. 2 MCMC of 10000 iterations have been computed. The aim is to assess whether the posterior distribution of β can capture the true value, that in this case is 2. Results with $C = c(X'X)^{-1},\ c = 1$ show that all the β parameters fluctuate around 0, so this prior is not suitable to represent the real value of parameter. Moreover, the sensitivity is 67.35% with a specificity fixed to 90%.

On the other hand, $C = cI,\ c = 1$ is fixed, and β parameters seem to be able to well represent the true value 2. A graphical check is shown in figure 5.1 where it is shown that β fluctuate around 2. Summaries of each β are collected in the table below. Sensitivity computed on 1 repetition results equal to 71.73 %.

β	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\beta_{1,1}$	1.878	1.996	2.029	2.030	2.063	2.184
$\beta_{2,1}$	1.770	1.910	1.944	1.943	1.976	2.139
$\beta_{1,2}$	1.878	2.007	2.045	2.045	2.084	2.249
$\beta_{2,2}$	1.813	1.948	1.990	1.988	2.025	2.163
$\beta_{1,3}$	1.896	2.012	2.046	2.046	2.079	2.201
$\beta_{2,3}$	1.716	1.878	1.913	1.913	1.945	2.085

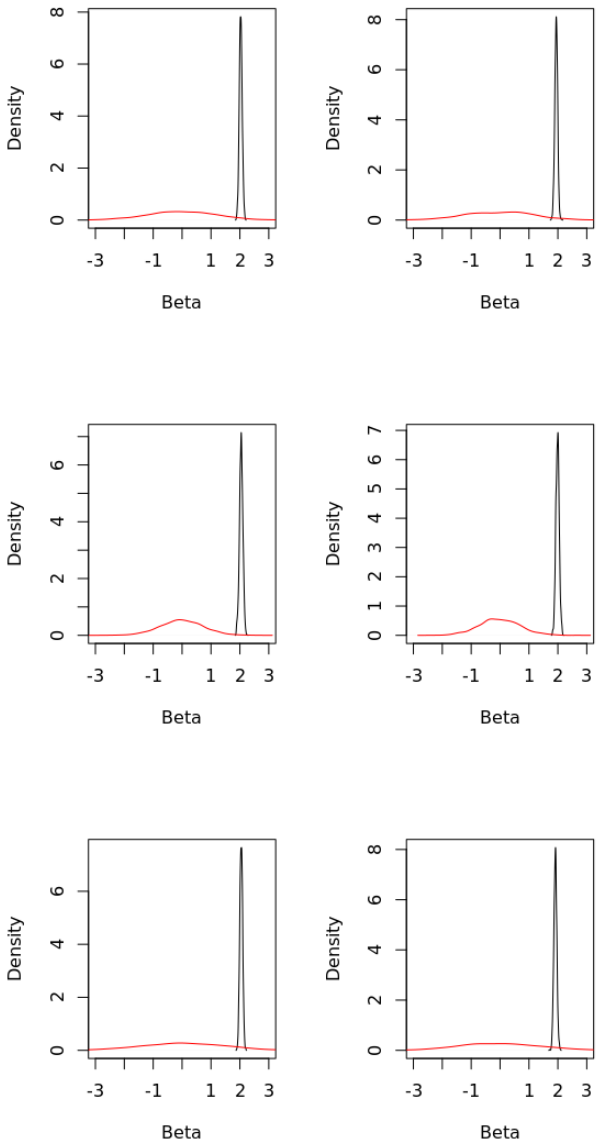

Figure 5.1: Posterior Beta vs Prior Beta

For the covariates-adjusted method matrix $C = I$ has been chosen to generate β covariance matrix.

3 scenarios are analyzed, whose results are reported in the next section. The scenarios consist in generating the biomarkers levels with 3 different values of β:

- $\beta = 0$
- $\beta = 1$
- $\beta = 2$

5.1.3 Repeated Simulations

For each scenario 10 simulated data-sets are analyzed. For each analysis we run 2 MCMC chains of 10000 iterations. Chains dimension is chosen with respect not to fall in lack of convergence. Some summary statistics or explanatory plots have been computed to assess the model fitting. Then, in the detection phase, the risk of disease has been computed on the 400 patients with respect to the updated data.

Model Fitting

Model fitting is assessed for the new method. Graphical comparisons between prior distribution and the posterior distribution of β parameters are reported for the first repetition, just as an example. Moreover, a graphical representation of biomarkers trajectories for 1 case patient and 1 control patient is included: real biomarkers levels are compared to predicted biomarkers level. The predicted biomarker level is computationally obtained in the following way:

- Control patients:
  1. the biomarker level is generated $Y = \theta + \beta X$;
  2. step (1.) is repeated for all the 10000 iterations;
  3. a mean is computed.
- Case patients:
  1. the time of the visit "t" is fixed;
  2. the biomarker level is generated $Y = \theta + \beta X + \gamma(t - \tau)^{(+)}$;
  3. step (1.) is repeated for all the 10000 iterations;
  4. a mean is computed;
  5. the process (1.)-(4.) is repeated for all the visits times "t".

On all the 10 repetitions random samples have been obtained for: the parameters posterior mean and the parameters Gelman-Rubin statistics $\hat{R}$. In this last phase Frequentist approach is applied to the results from Bayesian method applied in the current work.

Frequentist and Bayesian encounter in order to assure the convergence of chains and the accuracy of the obtained results.

True value $\beta = 0$

Plot of comparison between priors and posteriors of all the 6 β parameters is reported in figure 5.2. As it can be noted in the plots, all β are centred on their real value fixed in the generated data. Summaries of β parameters are reported in the table below.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\beta_{1,1}$	-0.133531	-0.003501	0.026977	0.029481	0.062053	0.193835
$\beta_{2,1}$	-0.20202	-0.08925	-0.05546	-0.05567	-0.02314	0.08397
$\beta_{1,2}$	-0.14568	0.00872	0.04842	0.04848	0.09136	0.23664
$\beta_{2,2}$	-0.208258	-0.043984	-0.003668	-0.005011	0.031397	0.228276
$\beta_{1,3}$	-0.14005	0.01535	0.04984	0.05108	0.08458	0.21568
$\beta_{2,3}$	-0.26356	-0.11583	-0.08075	-0.08205	-0.04966	0.06599

All of the parameters fluctuate around the their real value 0.

Plot of actual fitting of the model on real data, especially on a control patient, is reported in figure 5.3. Plots represent the predicted level of the 3 biomarkers and the real measurements at each visit for the control patient with ID=20. The same is done for one case patient. On the biomarkers trajectories a changepoint can be noticed. The plots are in figure 5.4. Real biomarkers level are represented by dots, while black line represents the predicted level and black dashed lines represent the 95% confidence interval. Red dashed line represents the level the biomarker would have had without the changepoint (it is made in order to appreciate the changepoint). Indeed, it is shown that the changepoint is caught in the first and in the third biomarker trajectory.

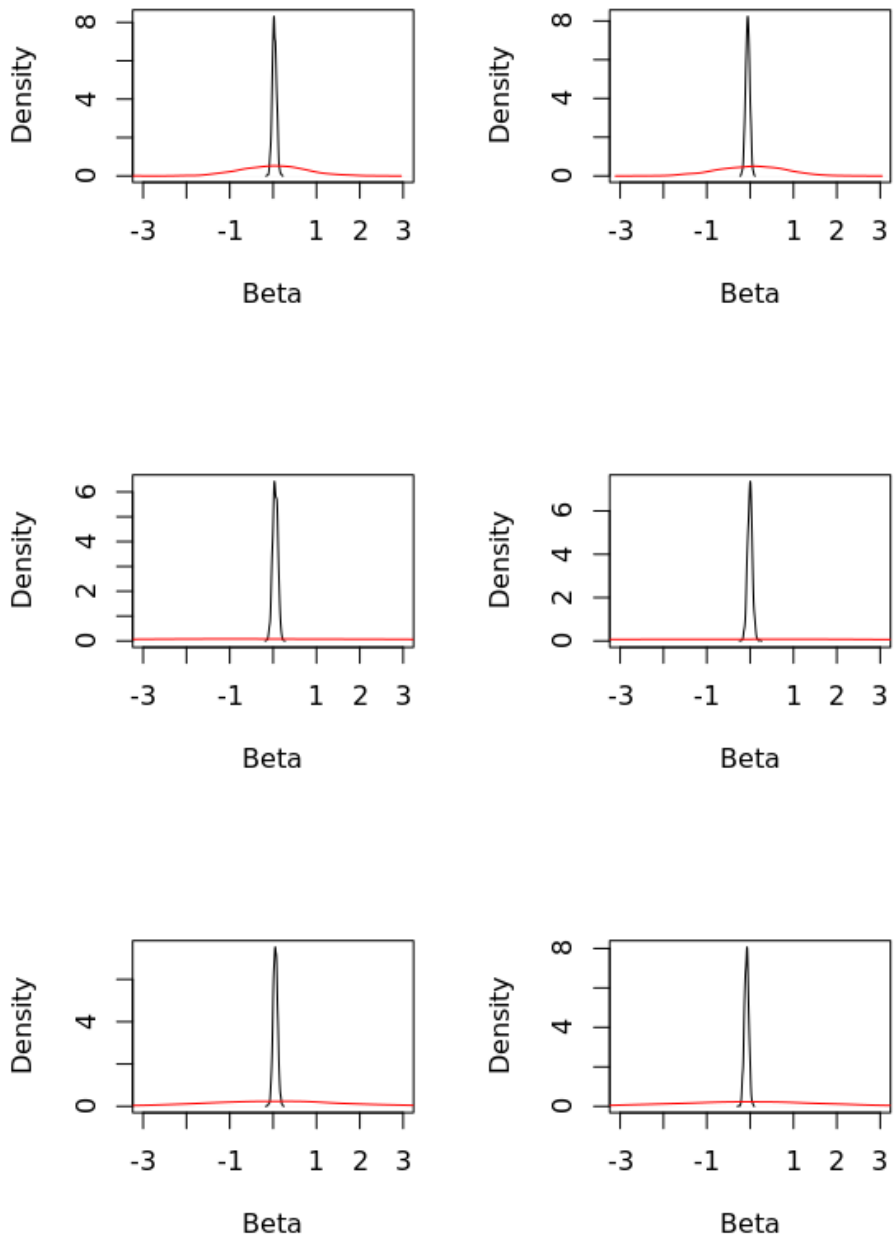

Figure 5.2: Prior vs Posterior

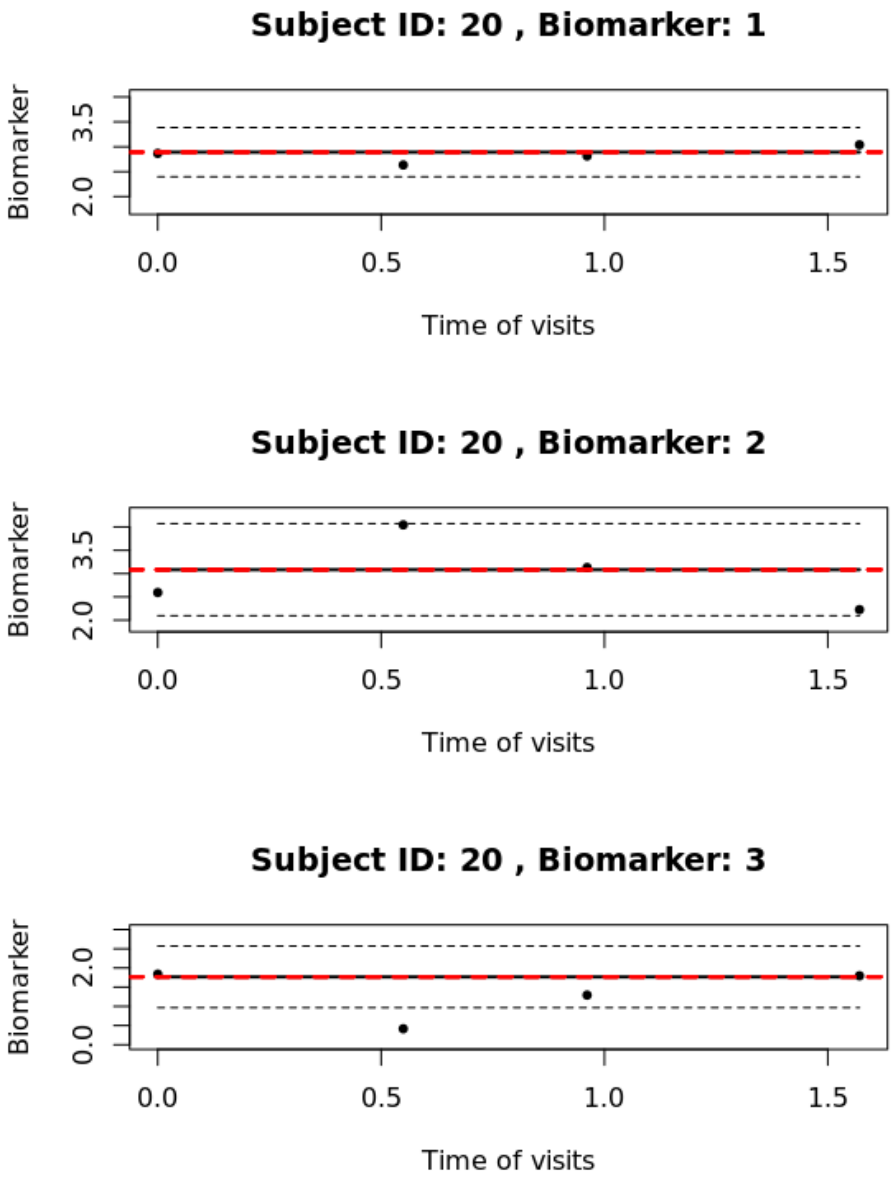

Figure 5.3: Control fitting

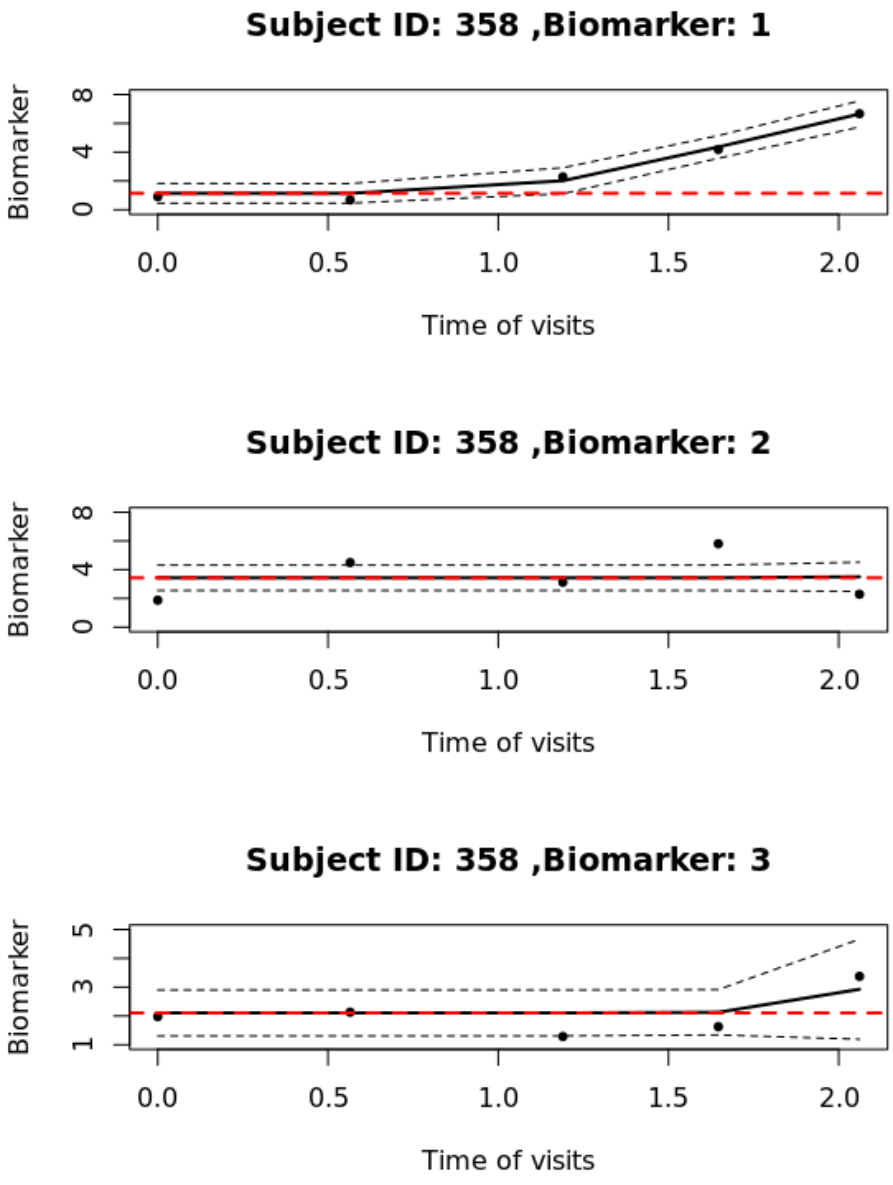

Figure 5.4: Case fitting

Within 10 repetitions for each parameter, parameters mean of the means have been collected and are compared with their true value, in the table below. Gelman-Rubin statistics stays under 1.02 in mean between repetitions for all the parameters. It can be concluded that there is not evidence for lack of convergence. Moreover, the mean value is close to the real value with whom each parameter has been generated.

Par.	True v.	Mean	$\hat{R}$
μ_{θ_1}	2.43	2.4156	1.0009
μ_{θ_2}	3.10	3.0959	1.0005
μ_{θ_3}	2.75	2.6723	0.9999
$\sigma^2_{\theta_1}$	0.79	0.8169	0.9998
$\sigma^2_{\theta_2}$	0.80	0.8770	1.0001
$\sigma^2_{\theta_3}$	0.79	0.7356	0.9996
σ^2_1	0.23	0.2416	0.9998
σ^2_2	1.35	1.3748	1.0011
σ^2_3	0.80	0.8178	1.0004
μ_I	0.15	0.3980	1.0011
η_I	0.1	0.1005	1.0023
μ_{γ_1}	1.87	2.3848	1.0290
μ_{γ_2}	1.92	1.9511	1.0267

Par.	True v.	Mean	$\hat{R}$
μ_{γ_3}	1.00	1.9462	1.0278
$\sigma^2_{\gamma_1}$	1.61	0.3892	1.0263
$\sigma^2_{\gamma_2}$	0.05	0.0293	1.0055
$\sigma^2_{\gamma_3}$	0.20	0.0444	1.0078
μ_{τ_1}	1.05	0.5366	1.0156
μ_{τ_2}	0.56	0.6260	1.0073
μ_{τ_3}	0.75	0.3972	1.0033
$\sigma^2_{\tau_1}$	0.82	0.6231	1.0035
$\sigma^2_{\tau_2}$	0.58	0.6227	1.0021
$\sigma^2_{\tau_3}$	0.70	0.5627	1.0004
$\sigma^2_{\beta_1}$	-	0.5067	1.0002
$\sigma^2_{\beta_2}$	-	0.5063	0.9997
$\sigma^2_{\beta_3}$	-	0.5003	0.9995

True value $\beta = 1$

The plot of comparison between priors and posteriors of all the 6 β is reported in figure 5.5. As it can be noticed in plots, all β are centred on their real value from the generated data while the prior in red is totally uninformative. Summaries of β parameters are reported as well in the table below. All the parameters fluctuate around their real value 1.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\beta_{1,1}$	0.8839	0.9959	1.0267	1.0275	1.0577	1.2041
$\beta_{2,1}$	0.7370	0.9085	0.9412	0.9408	0.9749	1.0894
$\beta_{1,2}$	0.842	1.008	1.048	1.047	1.087	1.233
$\beta_{2,2}$	0.8237	0.9485	0.9882	0.9878	1.0288	1.1872
$\beta_{1,3}$	0.9045	1.0094	1.0418	1.0441	1.0801	1.2622
$\beta_{2,3}$	0.7420	0.8786	0.9126	0.9129	0.9469	1.1028

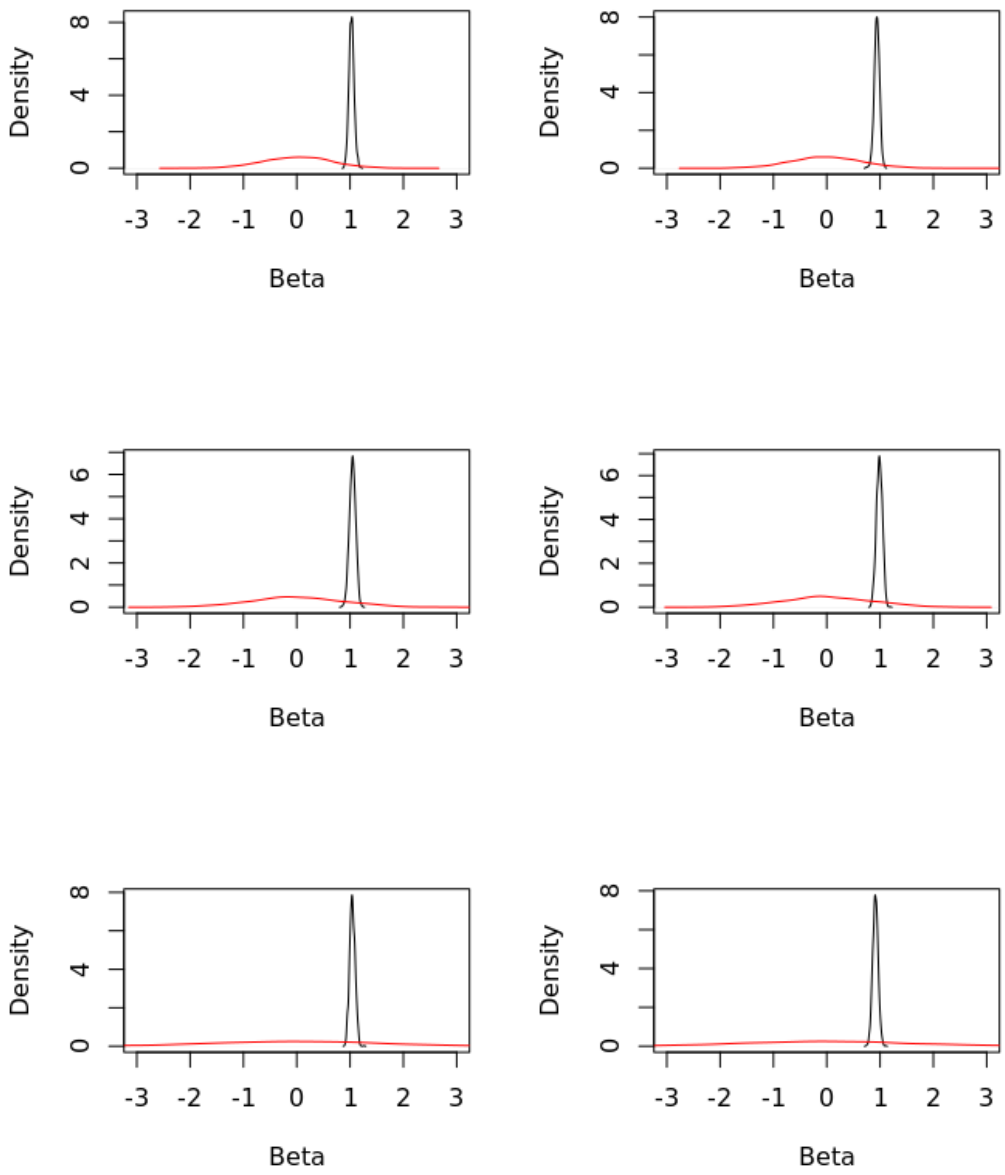

Figure 5.5: Prior vs Posterior

The plot of the actual fitting of the model on real data for a control patient, is reported in figure 5.6. Plots represent the predicted level of the 3 biomarkers (red line) and the real measurements at each visit (dots) for the control patient number 20.

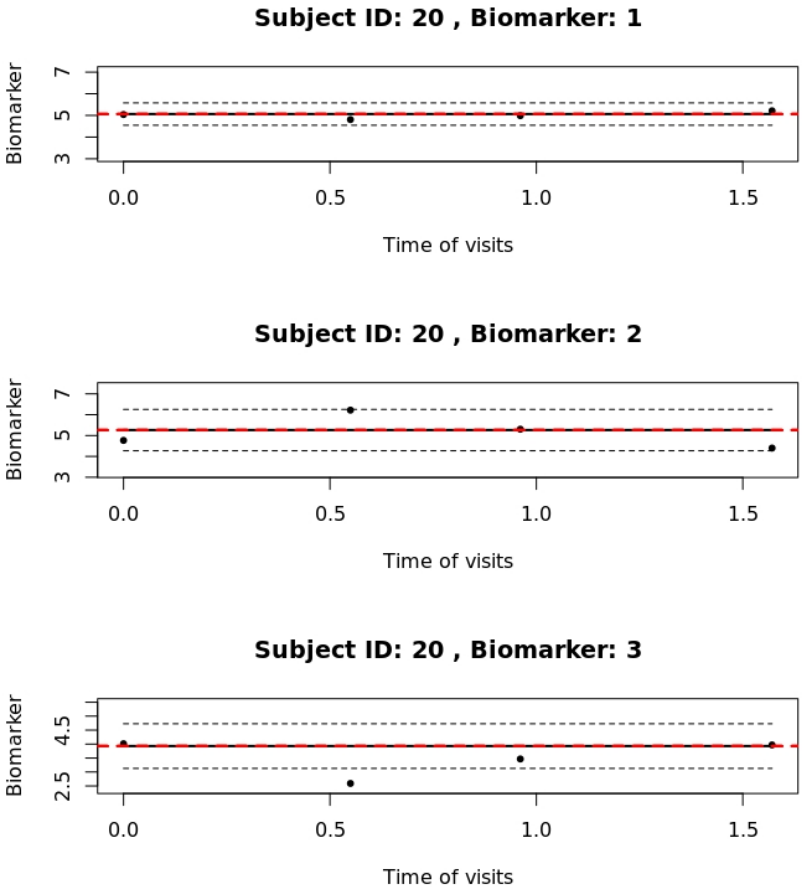

Figure 5.6: Fitting of the model on a control patient

The same is done for a case patient. On the biomarkers trajectories a changepoint can be seen. The plots are in figure 5.7. Within 10 repetitions for

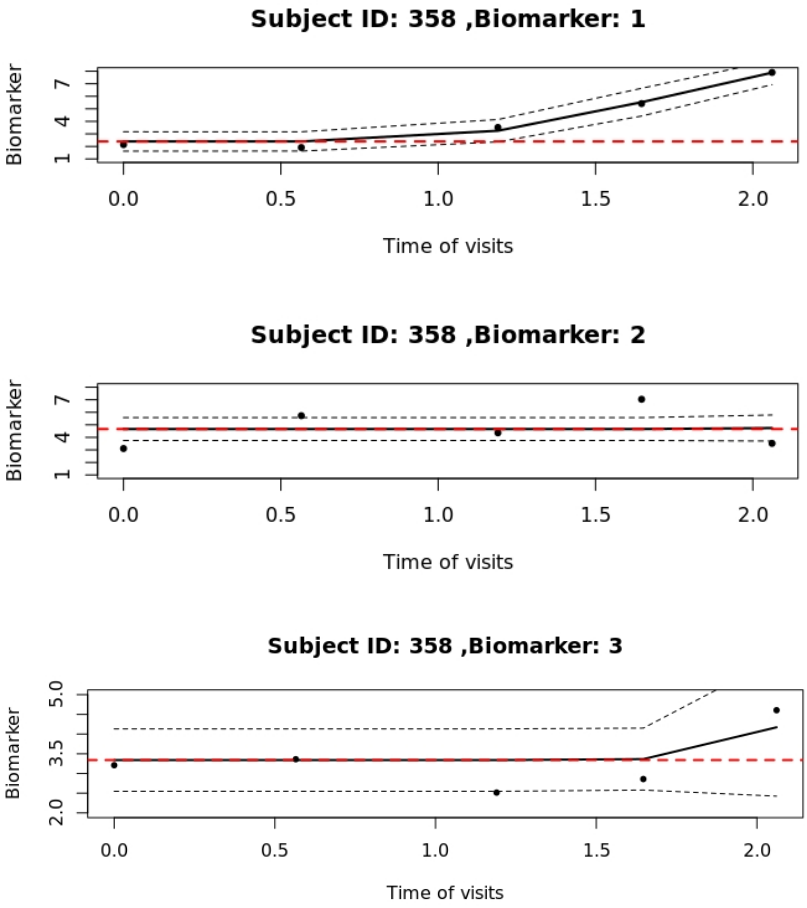

Figure 5.7: Fitting of the model on a case patient

each parameter, parameters mean of the means have been collected and are compared with their true value, in the table below. Gelman-Rubin statistic stays under 1.04 in mean between repetitions for all the parameters. It can be concluded that there is not evidence for lack of convergence. Moreover, the mean values are close to the real values with whome the parameters have been generated.

Par.	True v.	Mean	$\hat{R}$
μ_{θ_1}	2.43	2.4168	0.9999
μ_{θ_2}	3.10	3.0958	0.9996
μ_{θ_3}	2.75	2.6722	1
$\sigma^2_{\theta_1}$	0.79	0.8161	0.9999
$\sigma^2_{\theta_2}$	0.80	0.8767	0.9994
$\sigma^2_{\theta_3}$	0.79	0.7347	0.9998
σ^2_1	0.23	0.2417	1.0005
σ^2_2	1.35	1.3731	1.0014
σ^2_3	0.80	0.818	0.9996
μ_I	0.15	0.4049	1.0016
η_I	0.1	0.0989	1.0038
μ_{γ_1}	1.87	2.3985	1.0326
μ_{γ_2}	1.92	1.9587	1.0229

Par.	True v.	Mean	$\hat{R}$
μ_{γ_3}	1.00	1.9363	1.0091
$\sigma^2_{\gamma_1}$	1.61	0.3689	1.0183
$\sigma^2_{\gamma_2}$	0.05	0.0291	1.004
$\sigma^2_{\gamma_3}$	0.20	0.0456	1.0095
μ_{τ_1}	1.05	0.5184	1.0164
μ_{τ_2}	0.56	0.6129	1.0049
μ_{τ_3}	0.75	0.4052	1.0025
$\sigma^2_{\tau_1}$	0.82	0.619	1.0034
$\sigma^2_{\tau_2}$	0.58	0.617	1.0019
$\sigma^2_{\tau_3}$	0.70	0.5654	1.0011
$\sigma^2_{\beta_1}$	-	0.9827	1.0003
$\sigma^2_{\beta_2}$	-	1.0202	0.9999
$\sigma^2_{\beta_3}$	-	0.9874	0.9998

True value $\beta = 2$

Plot of comparison between priors and posteriors of all the 6 β is reported in figure 5.5. As it can be noticed in the plots, all β parameters are centred on their real value from the generated data while the prior in red is totally uninformative. Summaries of β parameters are reported as well in the table below. All parameters fluctuate around their real value 2.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\beta_{1,1}$	1.881	1.990	2.025	2.024	2.055	2.160
$\beta_{2,1}$	1.787	1.911	1.943	1.944	1.978	2.153
$\beta_{1,2}$	1.853	2.008	2.048	2.048	2.088	2.250
$\beta_{2,2}$	1.801	1.950	1.986	1.987	2.023	2.201
$\beta_{1,3}$	1.903	2.008	2.044	2.044	2.077	2.215
$\beta_{2,3}$	1.735	1.873	1.910	1.910	1.945	2.077

The plot of the actual fitting of the model on real data for a control patient, is reported in figure 5.6. The plots represent the predicted level of the 3 biomarkers (in red) and the real measurements (dots) at each visit for the control patient with ID=20.

The same is done for one case patient. On the biomarkers trajectories a changepoint can be noticed. The plots are in figure 5.10.

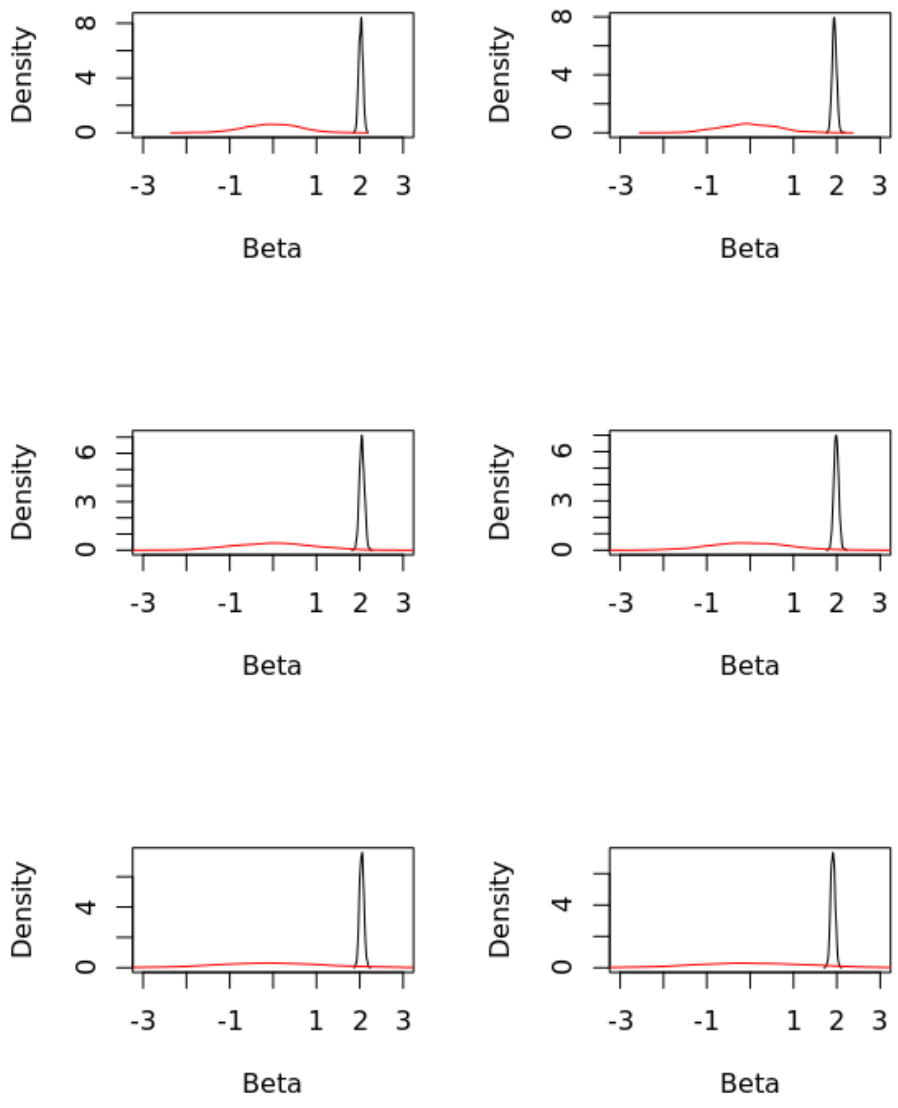

Figure 5.8: Prior vs Posterior

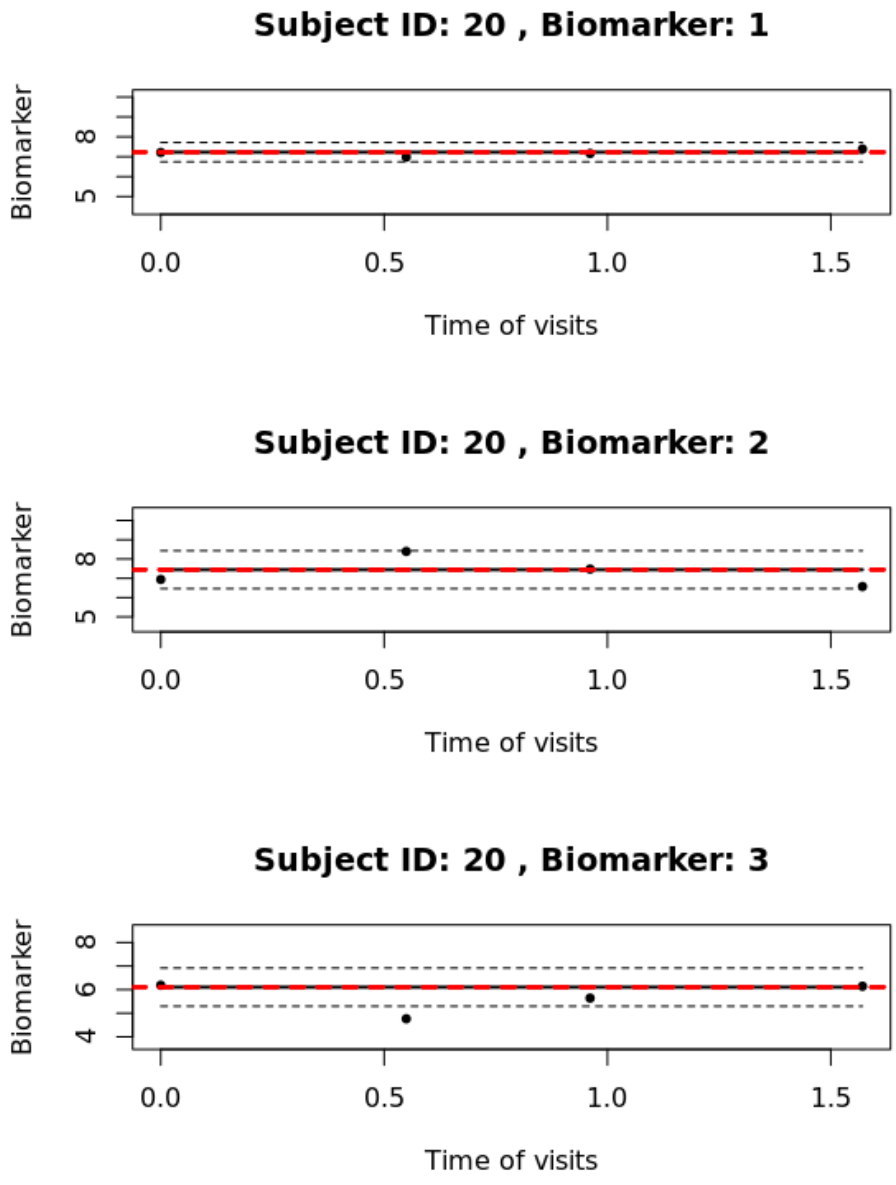

Figure 5.9: Fitting of the model on a control patient

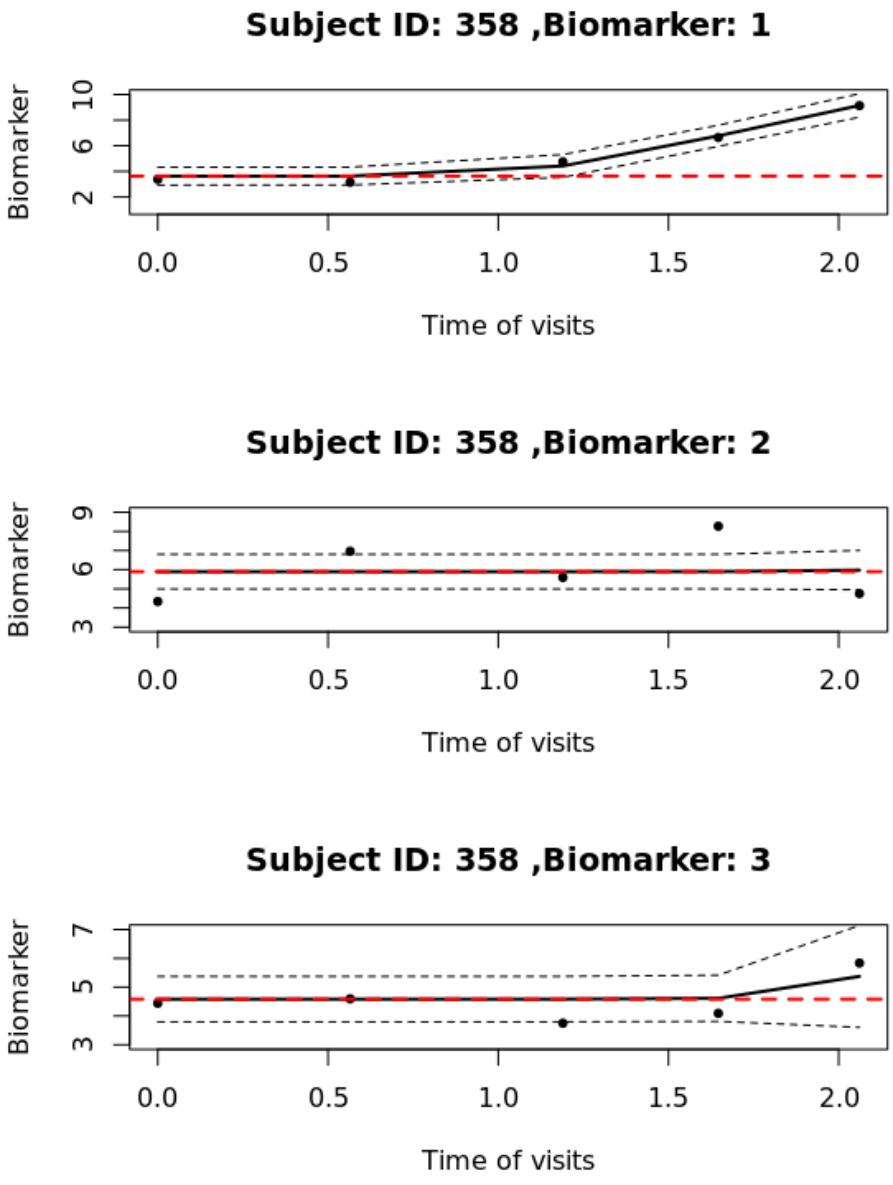

Figure 5.10: Fitting of the model on a case patient

Within 10 repetitions parameters mean of the means have been collected and are compared with their true value, in the table below. The Gelman-Rubin statistic stays under 1.04 in mean between the repetitions for all the parameters. It can be concluded that there is not evidence for lack of convergence. Moreover, the mean values are close to the real values with whom the parameters have been generated.

Par.	True v.	Mean	$\hat{R}$
μ_{θ_1}	2.43	2.4159	1.0001
μ_{θ_2}	3.10	3.0959	0.9998
μ_{θ_3}	2.75	2.6724	1.001
$\sigma^2_{\theta_1}$	0.79	0.816	1.0003
$\sigma^2_{\theta_2}$	0.80	0.8768	0.9999
$\sigma^2_{\theta_3}$	0.79	0.7352	1.0006
σ^2_1	0.23	0.2417	1.0006
σ^2_2	1.35	1.3741	1.0009
σ^2_3	0.80	0.8175	0.9998
μ_I	0.15	0.4004	1.0019
η_I	0.1	0.1003	1.0041
μ_{γ_1}	1.87	2.3967	1.0453
μ_{γ_2}	1.92	1.9604	1.0302

Par.	True v.	Mean	$\hat{R}$
μ_{γ_3}	1.00	1.946	1.026
$\sigma^2_{\gamma_1}$	1.61	0.3754	1.0214
$\sigma^2_{\gamma_2}$	0.05	0.0294	1.0048
$\sigma^2_{\gamma_3}$	0.20	0.0451	1.0095
μ_{τ_1}	1.05	0.5193	1.0155
μ_{τ_2}	0.56	0.6112	1.0064
μ_{τ_3}	0.75	0.3964	1.0037
$\sigma^2_{\tau_1}$	0.82	0.6196	1.0039
$\sigma^2_{\tau_2}$	0.58	0.6187	1.003
$\sigma^2_{\tau_3}$	0.70	0.5633	1.0023
$\sigma^2_{\beta_1}$	-	2.4447	1.0008
$\sigma^2_{\beta_2}$	-	2.5023	1.0003
$\sigma^2_{\beta_3}$	-	2.4846	0.9999

Detection

HCC detection is carried out with covariate-free method and with covariate-adjusted method. The accuracy in detection is compared between the 2 methods: indeed, fixing the specificity to 90%, the aim is computing the respective sensitivity. Relaxing the constraint on specificity to assume a unique value, for each repetition a ROC curve is computed in comparison between the new covariate-asjusted method and mFB approach.

Covariate-adjusted method with $\beta = 0$

Sensitivity with respect to a specificity of 90% within 10 repetitions are collected in the following table.

Simulation	Covariate model	Covariate-free model
1	71.43	75.51
2	75.51	77.55
3	79.59	81.63
4	71.43	77.55
5	77.55	73.47
6	73.47	85.71
7	77.55	75.51
8	75.51	71.43
9	81.63	79.59
10	77.55	77.55
Mean	76.12	77.55
S.e.	1.0556	1.2902

Sensitivity at the same specificity seems better in mFB method since all the covariates are in this simulation study not informative at all.

Then, the specificity is not fixed and the ROC curve is computed. It is represented for each repetition in figures 5.11, 5.12 in comparison between the covariate-adjusted method and the mFB. No big differences can be appreciated between plots from the 2 approaches.

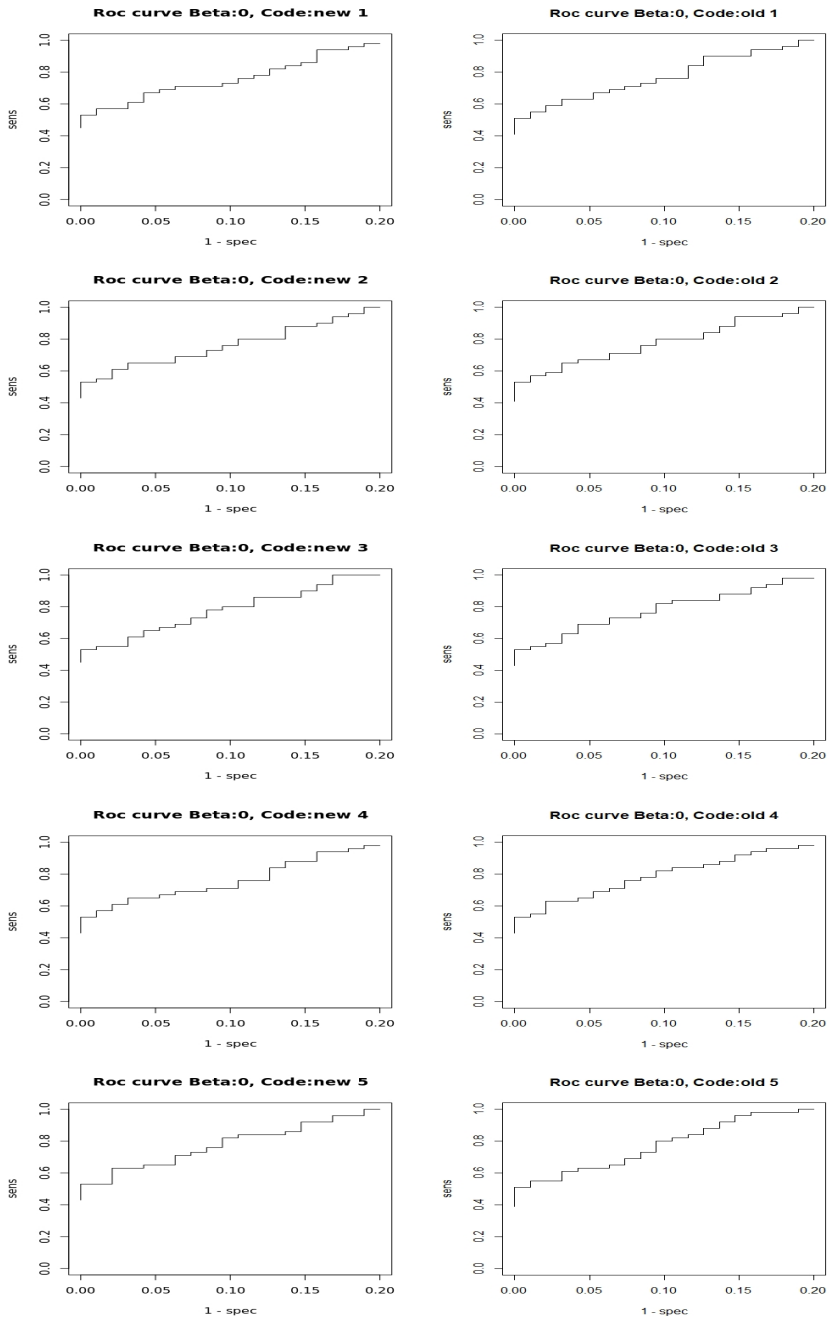

Figure 5.11: from 1 to 5 repetitions sensitivity comparison

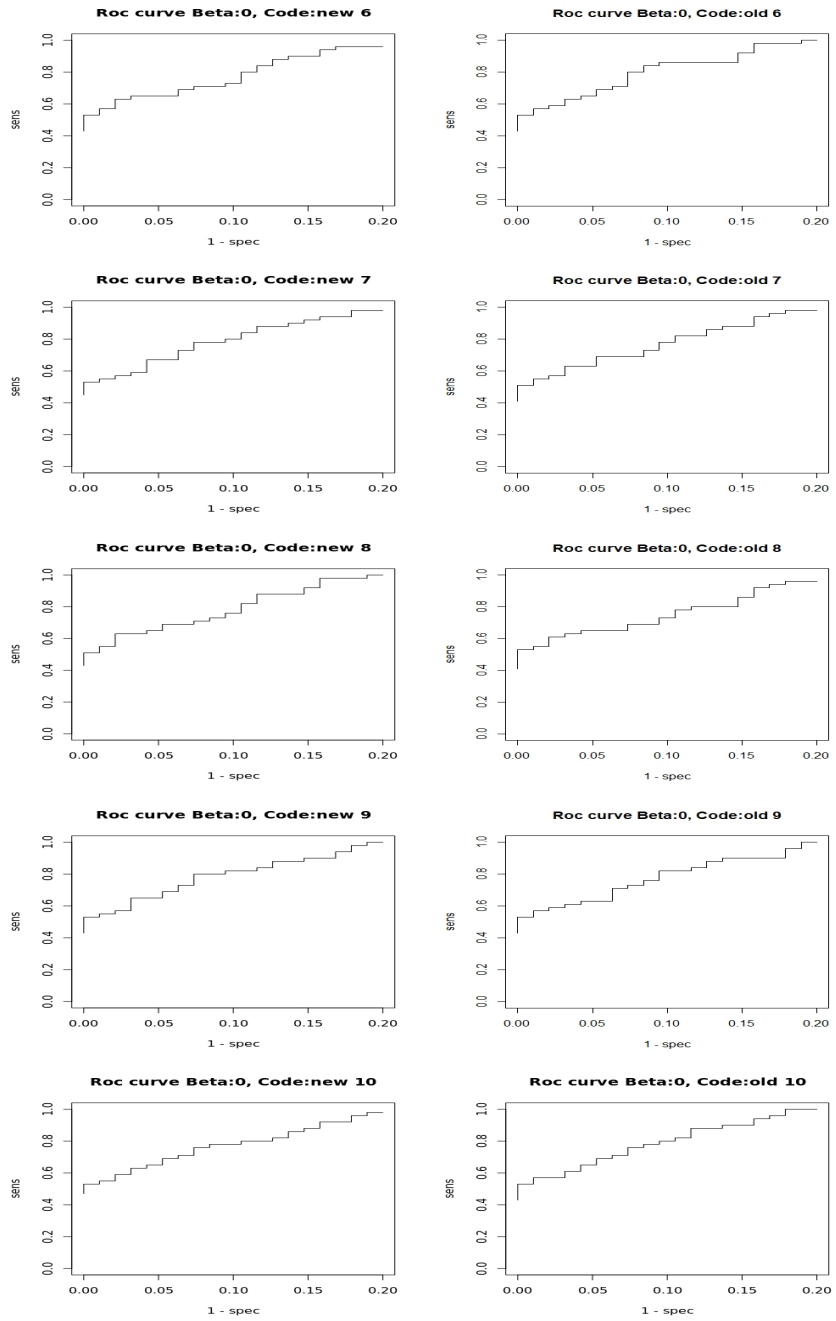

Figure 5.12: from 6 to 10 repetitions sensitivity comparison

Covariate-adjusted method with $\beta = 1$

The sensitivity with respect to a specificity of 90% within all 10 repetitions are collected in the following table.

Simulation	Covariate model	Covariate-free model
1	73.47	69.39
2	73.47	69.39
3	71.43	77.55
4	81.63	75.51
5	73.47	71.43
6	73.47	69.39
7	73.47	71.43
8	79.59	63.27
9	73.47	65.31
10	71.43	73.47
Mean	74.49	70.61
S.e.	1.0644	1.3699

Sensitivity at the same specificity seems better in the new covariate-adjusted method since all covariates are informative.

Then, the specificity is not fixed and the ROC curve is computed. ROC curves are represented for each repetition in figures 5.13, 5.14 in comparison between the covariate-adjusted method and the mFB.

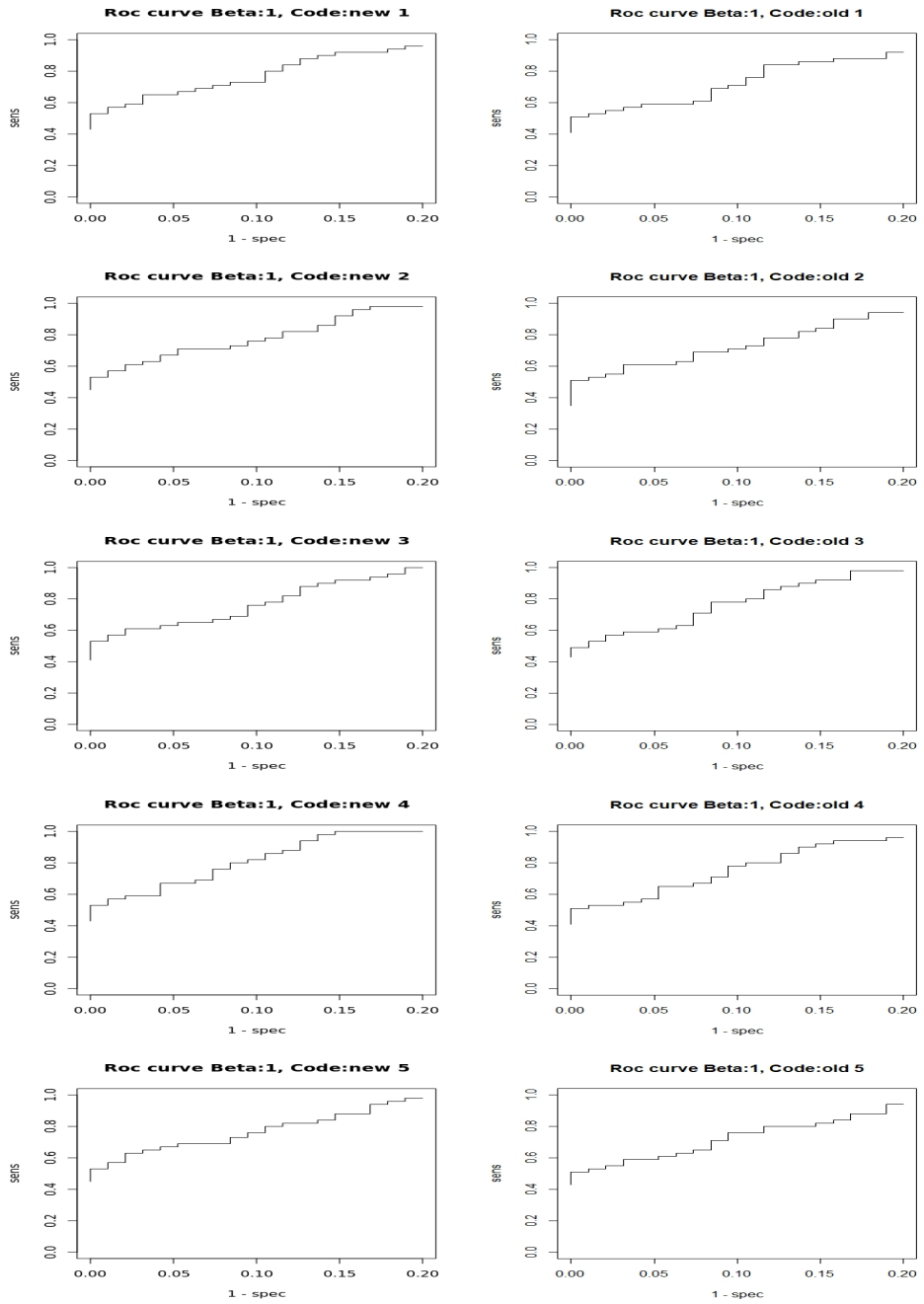

Figure 5.13: from 1 to 5 repetitions sensitivity comparison

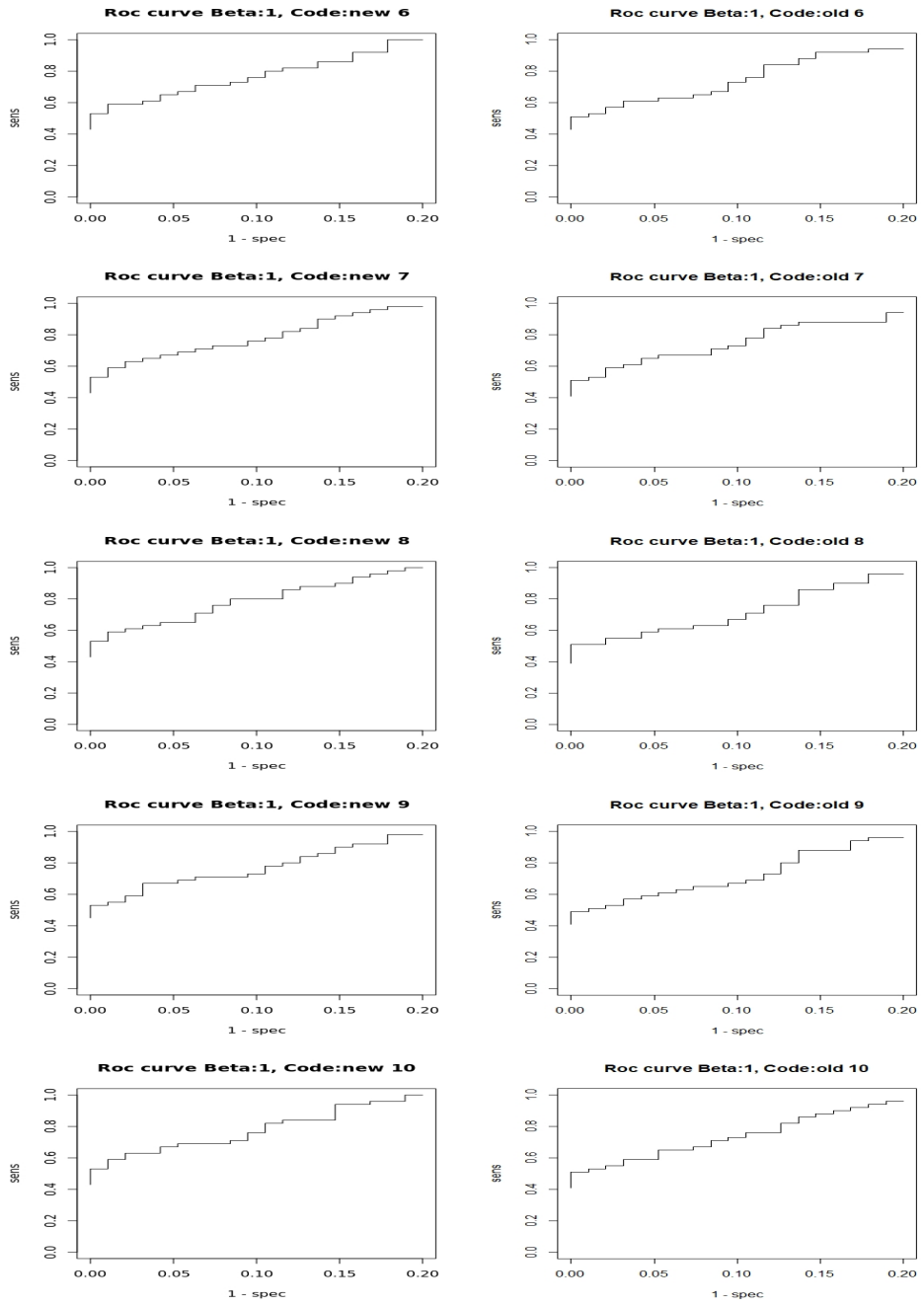

Figure 5.14: from 6 to 10 repetitions sensitivity comparison

Covariate-adjusted method with $\beta = 2$

Sensitivity with respect to a specificity of 90% within all 10 repetitions are collected in the following table.

Simulation	Covariate model	Covariate-free model
1	73.47	61.22
2	73.47	69.39
3	77.55	65.3
4	73.47	67.35
5	73.47	75.51
6	77.55	67.35
7	77.55	63.27
8	77.55	75.51
9	75.51	75.51
10	79.59	69.39
Mean	75.918	68.981
S.e.	0.7323	1.6325

Sensitivity at the same specificity seems better in the new covariate-adjusted method since all the covariates are informative.

Then, the specificity is not fixed and the ROC curve is computed. ROC curves are represented for each repetition in the figures 5.15, 5.16 in comparison between the new method and the old one.

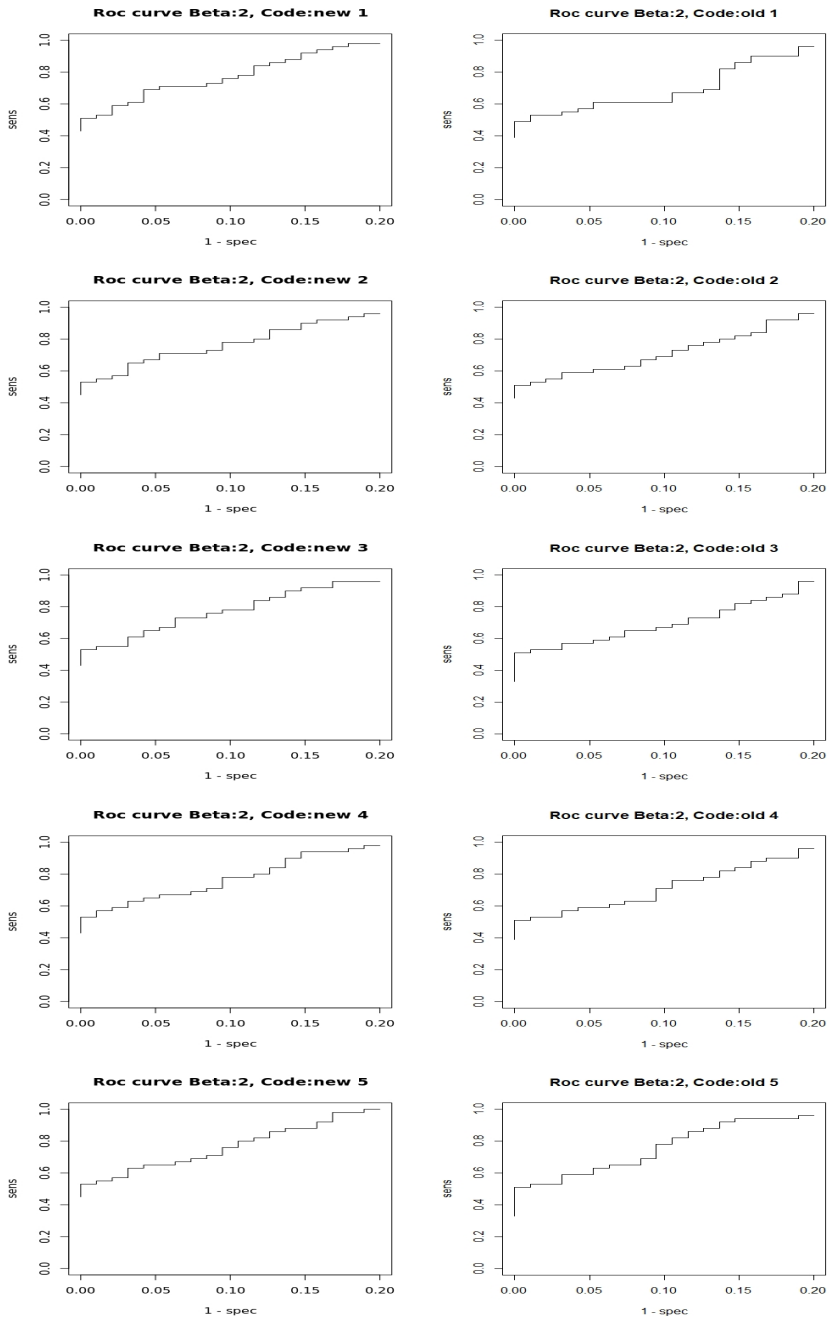

Figure 5.15: from 1 to 5 repetitions sensitivity comparison

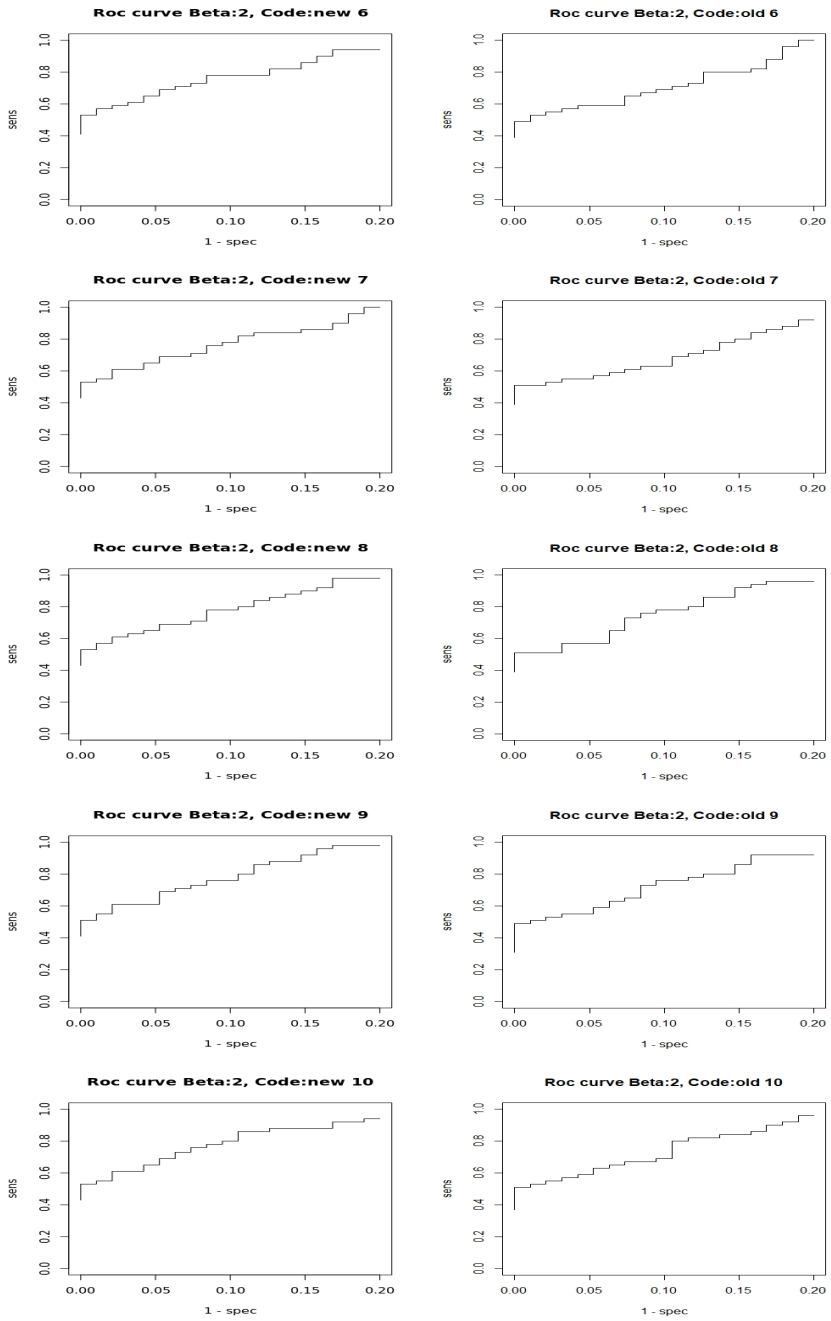

Figure 5.16: from 6 to 10 repetitions sensitivity comparison

5.2 HALT-C trial

The accuracy of the screening algorithm has been evaluated on the simulated data, therefore it can be now evaluated in cirrhosis patients from the HALT-C trial. Trajectories of log(AFP) and log(DCP+1) are assumed to follow the joint model described above [Tayob et al., 2018]. The logarithmic transformation has been made on biomarkers to keep low their range of variability and make them more symmetric.

Given the absence of a validation data-set, computation of posterior risk and consequently sensitivity are in-sample. That is the reason why results will be optimistic, and this fact has to be taken into account.

5.2.1 Data descriptive summaries

The Halt-C trial data carries a big amount of information. Patients are in total 409. The dataset consists of:

- RAND GRP indicates the randomization group (1=treatment, 2=control). 197 patients are assigned to treatment, 212 to control. Treatment consists of an interferon-based therapy.
- CIRRHOSIS, indicates if patients have cirrhosis (1=yes, 0=no). The 100% of HALT-C patients have cirrhosis.
- HCC, indicates if patients have HCC (1=yes, 0=no). Cotrols are 361, cases are 48.
- days HCC counts how many days before HCC has be detected.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	1852.00	2289.00	2186.28	2590.00	3161.00

- Early stage HCC indicates if HCC is detected at early stage (1=yes, 0=no). 36 are early detected out of 48 cases.
- VISIT DAYS indicate days from randomization to visit

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-965.00	-252.00	-217.00	-231.86	-183.00	-22.00

- AFP level at each visit

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.30	5.90	10.10	23.60	19.60	7051.20	371.00

- DCP level at each visit

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	11.10	22.40	91.17	41.20	29022.90	1747.00

- AGE at enrollment in years

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.00	46.00	50.00	50.22	54.00	77.00

- FEMALE indicates the gender (0:male, 1:female). 112 patients are females. 297 are males.

- HISP, WHITE, BLACK, OTHERS are dichotomous and indicate if the patient is or Hispanic, white, black either none of those races. 292 are white, 62 are black, 49 are hispanics and 6 of other races.

- FIBRO ISHAK score indicates the liver fibrosis stage in which the patient is at screening biopsy. Stages are discrete and they can be divided in the following way: "non-significant fibrosis" for Ishak score = (0, 1, 2); "significant fibrosis" for Ishak score = (3, 4); "advanced stage of fibrosis" for Ishak score = (5, 6). An increase in severity of fibrosis is linear between stages [Rosenberg et al., 2004]. 224 patients have fibrosis, 185 have cirrhosis.

- alt, ast indicate results of blood tests named respectively Alanine Aminotransferase (ALT) and Aspartate Aminotransferase (AST) that check for liver damages [Wikipedia contributors, 2019a], [Wikipedia contributors, 2019b].

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.00	65.00	95.00	117.07	144.00	647.00

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.00	60.00	86.00	102.61	128.00	518.00

- DIABETES, indicates if patients have diabetes (0:not, 1:yes). 329 are positive, 80 are negative.
- BMI indicates body mass index at baseline (weight (kg)/height$(m)^2$)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
19.69	26.43	29.65	30.28	33.01	48.97

- EVERDRANK indicates if the patient has ever drank in the lifetime (0:not, 1:yes). 341 patients did, 68 did not.
- LIFE DRINKS indicates the total number of drinks lifetime

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	1596.00	8424.00	19141.34	25110.00	357314.40	2.00

- AVE GRAMS P D indicates the average grams of alcohol per day

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	2.54	11.72	26.41	33.30	434.79	2.00

- SMOKE CIG NOW indicates if the patient smokes. 120 used to smoke cigarettes at the enrollment time. 289 did not.
- SMOKESTAT indicates the amount of smoking (1=Never; 2=Not now; 3=Now less than 1pack/day; 4=Now $\geq$ 1pack/day)

1	2	3	4
91	198	70	50

- PACKYEARS indicates how many cigarettes packs do the patient smoke

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.00	0.90	8.00	14.35	23.00	92.00	6.00

- platelets, with unit of measurement x1000/mm3

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.00	91.00	125.00	133.81	163.00	425.00

- albumin, with unit of measurement g/dL

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.70	3.50	3.80	3.76	4.10	4.90

- tot bilirubi, with unit of measurement mg/dL

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.20	0.60	0.80	0.89	1.10	3.80

5.2.2 Covariate-adjusted approach on HALT-C

Some results of the new approach applied to real data from Halt-C Trial are reported in this section. First of all, results from both the old and the new approach are reported in order to make a comparison between methods.

Approach	Sensitivy (ROC 0.1)	AUC
Fixed cut-off AFP	60.42%	0.84
Fixed cut-off DCP	56.25%	0.78
mFB	89.58%	0.95
2 covariates-adjusted	83.33%	0.89
8 covariates-afjusted	85.42%	0.93

Some other statistics about 2 covariates-adjusted approach and 8 covariates-adjusted approach are reported in the following sections.

AGE and FEMALE

Summaries of β parameters with respect to age and female are reported in the following table. Subscript "1" indicates AFP, subscript "2" indicates DCP.

Par.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\beta_{AGE,1}$	-0.013841	-0.008650	-0.004570	-0.002062	0.001148	0.018818
$\beta_{F,1}$	0.02579	0.26322	0.32484	0.32002	0.38101	0.55157
$\beta_{AGE,2}$	-0.019894	-0.007863	-0.004939	-0.005083	-0.002342	0.008819
$\beta_{F,2}$	-0.82371	-0.47719	-0.40728	-0.40862	-0.34004	-0.03669

Plots of comparison between not informative priors (in red) and posteriors (in black) are in figure 5.17. Posterior distributions result more informative than priors. Female seems to have a positive effect on AFP trajectory and a negative effect on DCP trajectory. While age seems not significant but, since it is measured in years it has an appreciable effect only over time.

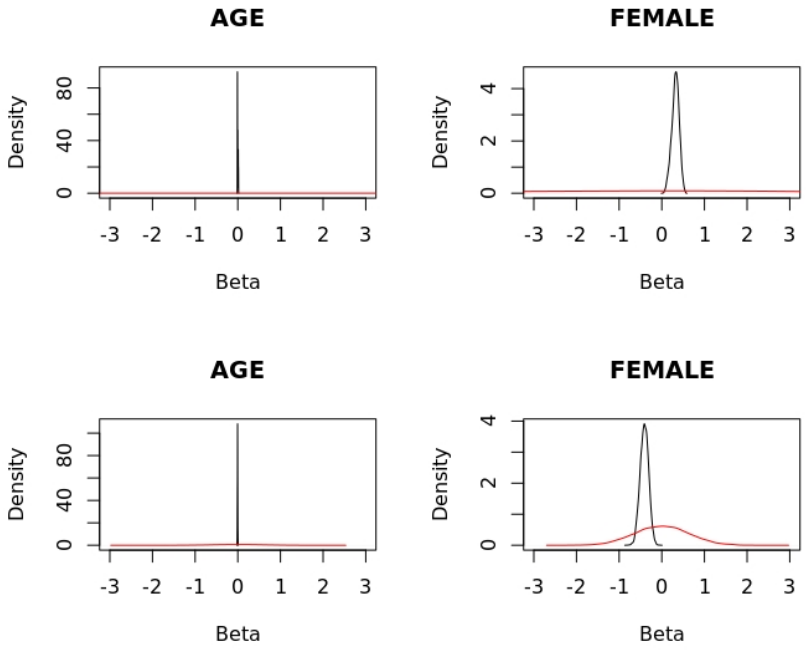

Figure 5.17: Prior vs Posterior

Some fitting graphical analysis are in figure 5.18. Four different case patients are analyzed in order to observe all the possible situations with 2 biomarkers AFP and DCP. Subject 1 has high probability of observing a changepoint in both biomarkers AFP and DCP, respectively 94% and 76%;

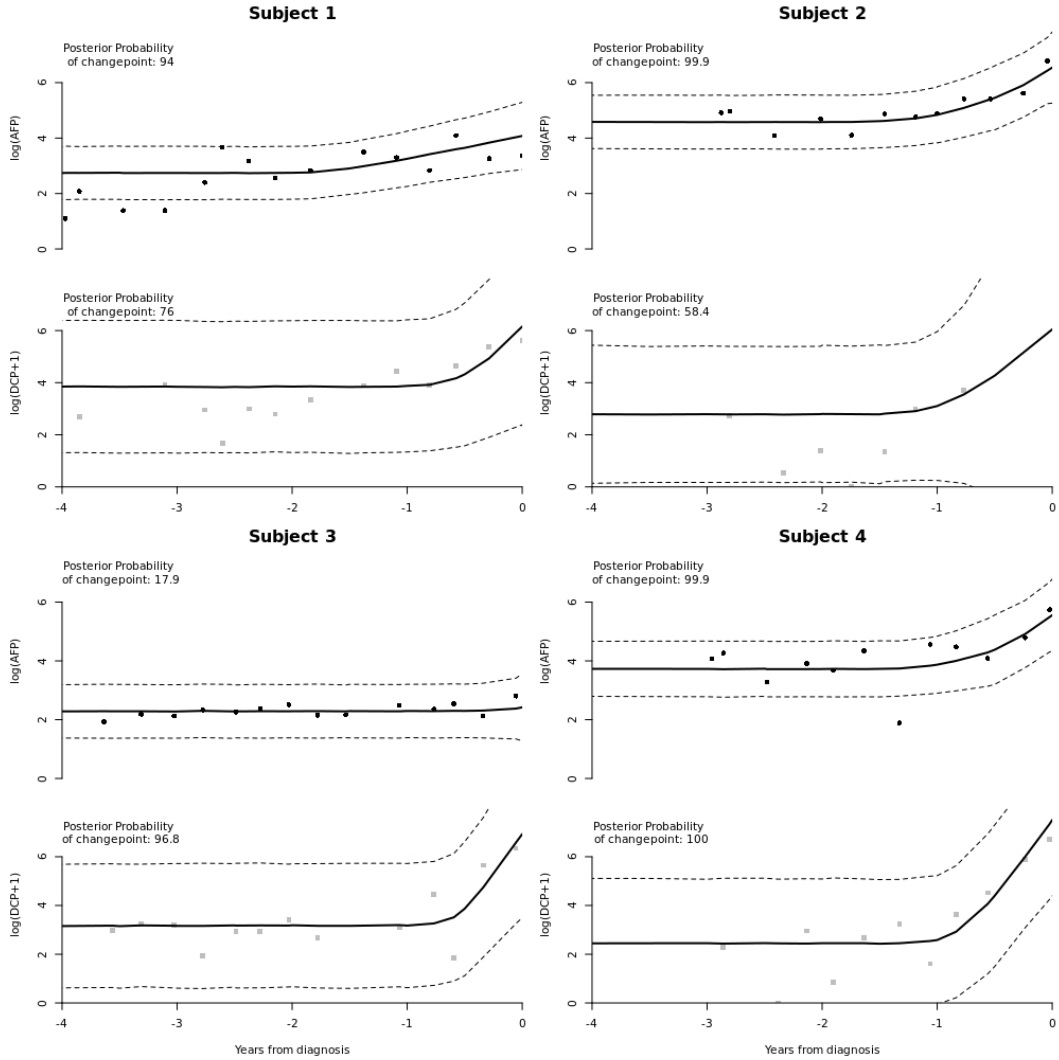

Figure 5.18: New Approach Fitting

subject 2 has an almost certain probability at 99.9% of observing a changepoint in AFP trajectory but a lower probability at 58.4% of observing a changepoint in DCP; subject 3 does not show a changepoint in AFP but it does in DCP with probability 96.8%; subject 4 has a very high probability of observing a changepoint in both AFP and DCP, respectively 99.9% and 100%. Timing of changepoints within patients seems to be similar except in subject 1.

Posterior values of used parameters are collected in the following table:

Parameter	Mean	Sd	$\hat{R}$
$\mu_{\theta_{AFP}}$	2.4681	0.4011	1.1248
$\mu_{\theta_{DCP}}$	3.4021	0.217	1.071
$\sigma^2_{\theta_{AFP}}$	0.8148	0.0596	1.0099
$\sigma^2_{\theta_{DCP}}$	0.6808	0.0573	0.9999
σ^2_{AFP}	0.2020	0.0037	1.0036
σ^2_{DCP}	1.3448	0.027	1.0006
μ_I	-0.1364	0.1998	1.0052
η_I	0.1174	0.0499	1.0015
$\mu_{\gamma_{AFP}}$	1.8898	0.2751	1.0006
$\mu_{\gamma_{DCP}}$	1.9059	0.2519	1.0432
$\sigma^2_{\gamma_{AFP}}$	1.5345	0.8566	1.0095
$\sigma^2_{\gamma_{DCP}}$	0.0555	0.0673	1.0133
$\mu_{\tau_{AFP}}$	1.0638	0.333	1.0118
$\mu_{\tau_{DCP}}$	0.6205	0.2807	1.0104
$\sigma^2_{\tau_{AFP}}$	0.7855	0.2701	1.0045
$\sigma^2_{\tau_{DCP}}$	0.5918	0.2061	1.0015
$\sigma^2_{\beta_{AFP}}$	0.5295	0.5115	1.0001
$\sigma^2_{\beta_{DCP}}$	0.5543	0.6343	1.0001

AGE, FEMALE, HISP, BLACK, OTHERS, FIBRO ISHAK, alt, ast

Combining 2 biomarkers AFP and DCP sensitivity increases significantly from 2 covariates model to 8 covariates model. The increase amounts to approximately 2 percentage points.

Some of the pre-clinical features have been selected because it was interesting to assess their effect on AFP and DCP trajectories. Selected variables or have a demographic nature either are about liver disease severity.

Summaries of β parameters related to selected covariates are reported in the following tables. They are referred to AFP, indexed as "biomarker 1" and DCP, indexed as "biomarker 2". Covariate AGE does not seem to be significant in both biomarkers but this phenomenon is only due to its measurement units taken in year; FEMALE is a dichotomous variable and represents the effect of female versus male; HISP, BLACK and OTHERS are 3 dummy variables and they represent the effect of ethnicity they are referred to with respect to the baseline ethnicity WHITE. FIBRO ISHAK (shorter ISHAK) represents the effect of cirrhosis (stage 6) with respect to fibrosis as baseline (stage 5); alt and ast represent the effect of blood tests they are referred to.

β_{AFP}	Min	1 qu.	Median	Mean	3 qu.	Max
AGE 1	-0.028077	-0.013311	-0.009654	-0.009442	-0.005733	0.006802
FEM 1	-0.1608	0.1271	0.1989	0.1896	0.2588	0.4818
HISP 1	-0.2432	0.1860	0.2719	0.2736	0.3632	0.6697
BLACK 1	0.08213	0.45539	0.52728	0.52358	0.59701	0.91460
OTHERS 1	-0.2318	0.4214	0.6239	0.6286	0.8230	1.8280
ISHAK 1	-0.10347	0.06982	0.10948	0.10780	0.15072	0.28528
alt 1	-0.006334	-0.003419	-0.002779	-0.002766	-0.002115	0.001614
ast 1	0.000709	0.004814	0.005698	0.005659	0.006535	0.009571

β_{DCP}	Min	1 qu.	Median	Mean	3 qu.	Max
AGE 2	-0.027529	-0.012547	-0.008719	-0.008850	-0.005136	0.010819
FEM 2	-0.80763	-0.53064	-0.46137	-0.46173	-0.39379	-0.06722
HISP 2	-0.439255	-0.081797	0.001645	0.003076	0.085616	0.434941
BLACK 2	-0.53304	-0.18956	-0.11363	-0.11477	-0.04048	0.37260
OTHERS 2	-0.73698	0.03823	0.22509	0.22221	0.40179	1.35373
ISHAK 2	-0.161891	0.008433	0.047827	0.047417	0.086879	0.251227
alt 2	-0.009463	-0.005945	-0.005257	-0.005253	-0.004567	-0.001343
ast 2	0.001682	0.006212	0.007032	0.007039	0.007889	0.011551

Model fitting is assessed graphically. Four different case patients are analyzed in order to observe all the possible situations with 2 biomarkers AFP and DCP. Subject 1 has high probability of observing a changepoint in

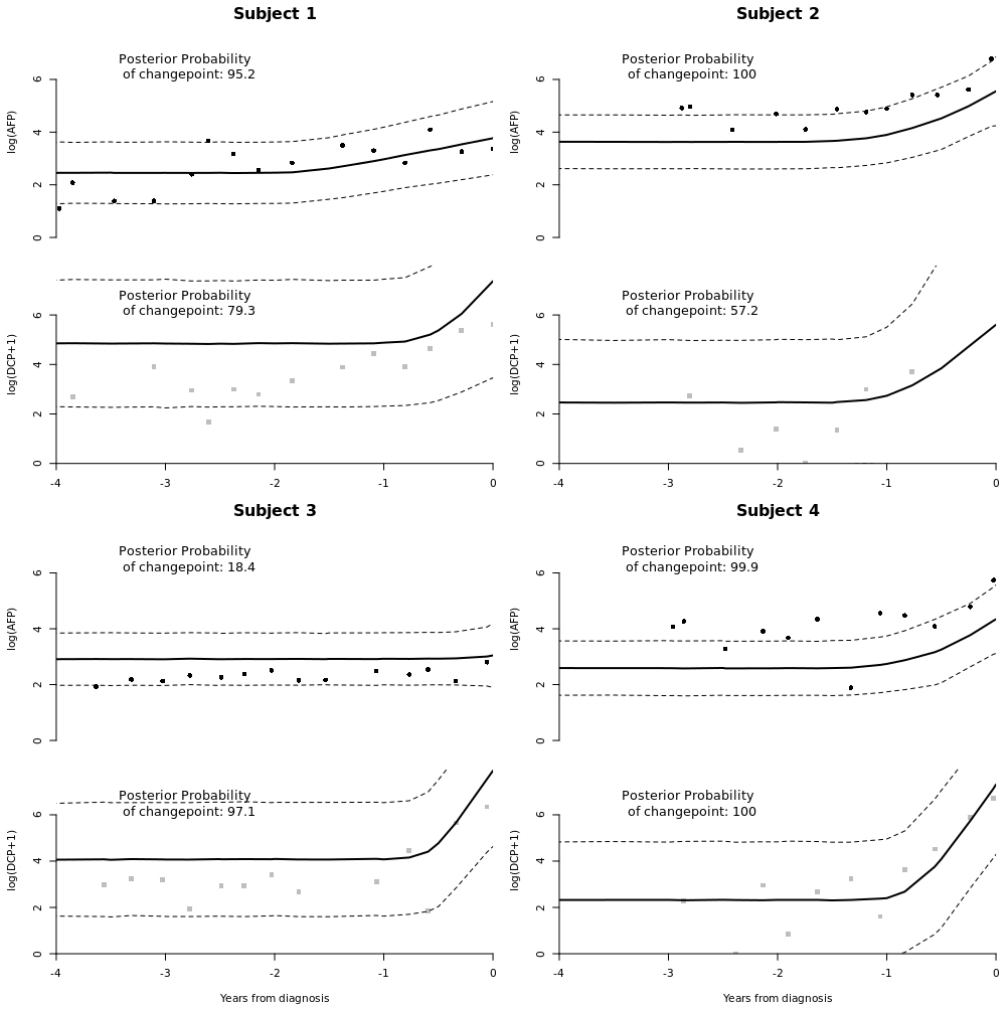

Figure 5.19: 8 covariates approach fitting

both biomarkers, 95.2% and 79.3% respectively for AFP and DCP; timing of the 2 changepoints seems to be vey different. Subject 2 has the certainty of observing a changepoint in AFP trajectory and a less clearly defined changepoint in DCP compared to AFP with probability 57.2%; subject 3

shows a not significant changepoint in AFP but a significant one in DCP with probability 97.1%; subject 4 has a very high probability of observing a changepoint in both AFP and DCP, respectively at 99.9% and 100%.

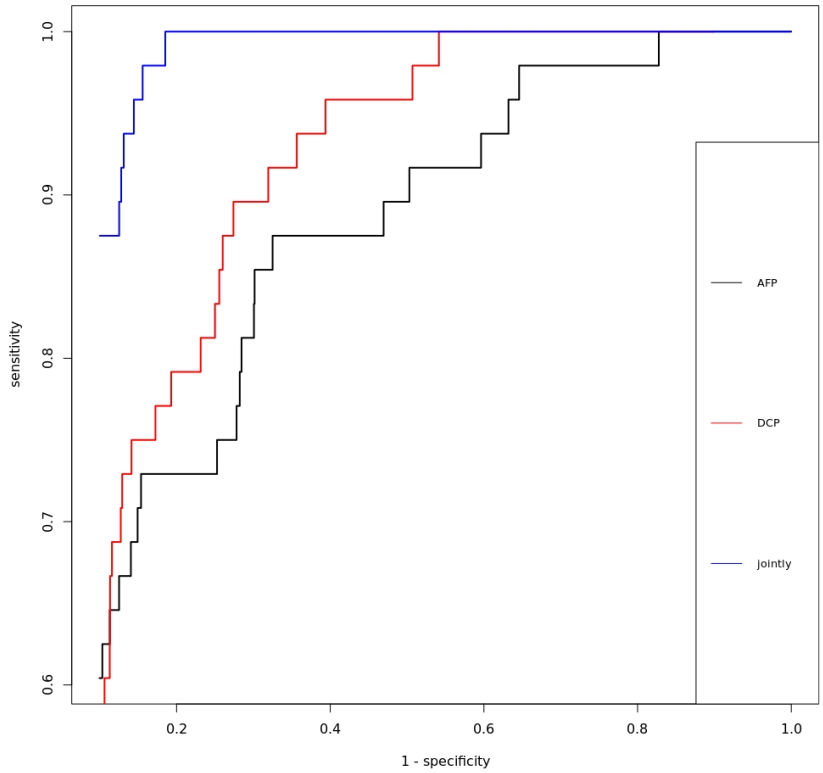

Figure 5.20: ROC curve on AFP, DCP, jointly

ROC curves are reported in figure 5.20, where AFP with a fixed cut-off model is represented in black, DCP with a fixed cut-off is in grey and model combining AFP and DCP with subject-specific cut-off is in blue. It is shown that the joint model has always the highest sensitivity, keeping the specificity fixed.

Chapter 6

Discussion

Early detection of HCC is an incredible powerful tool to allow patients to undertake successful treatments. Screening tests for early detection have to be as non-invasive as possible and as inexpensive as possible to allow its widespread use. Longitudinal biomarkers are a widely used diagnostic tool because they totally respect the desirable characteristics of a screening test. Biomarkers methods tend to have a high sensitivity but a low specificity, that is the reason why a priority in screening tests is to maintain low the false positive rate (FPR) in order to reduce costs and unnecessary anxiety. Indeed, it is not preferable to make patients undergo additional tests to ensure that their real disease status is the one predicted, if they are actually healthy. Multiple longitudinal biomarkers are necessary to obtain a screening test with both high sensitivity and high specificity [Pepe et al., 2001].

Usually survaillance programmes are taken on highly risk chirrosis patients, like patients in HALT-C trial. Patients are recommended to undertake ultrasonography along with measurements of AFP. Moreover, lately DCP showed its potential in higher sensitivity.

But AFP and DCP with fixed threshold for posterior risk computation turned out to have a lower sensitivity than jointly modeling their changepoints [Tayob et al., 2018]. Indeed, the aim with mFB model was to borrow information across biomarkers to identify changepoints that were more subtle. The same has been done in this work (some graphical examples are in sections 5.1.3, subject 3 in figure 5.18). In addition, we wanted to develop an extension of the mFB model to increasingly well capture the wide ranges of trajectories and identify future HCC cases earlier with more

accuracy.

It emerged that covariate-adjusted method proposed in this work is more likely to have higher sensitivity than mFB approach when prediction is made using a training set and a validation set (section 5.1.3). Indeed, the logic is including in the model covariates that may capture a component of the variation of the biomarker trajectory, and in that way improving the accuracy of the screening test.

On the contrary, results from Halt-C trial data are optimistic because the posterior risk computation is in-sample. If predictions are made in-sample then mFB is enough flexible to give very good results. Sensitivity from mFB results 89.5% as in Tayob et al. [2018]. It is higher than the sensitivity in model proposed here with 8 covariates that is about 85.42%.

Another reason to explain the lower sensitivity in covariate-adjusted approach may be that some variables have to be excluded from the model. Variables selection could be an crucial step to add to the algorithm in order to sharpen the accuracy of disease status prediction. This is a sketch for a future development. Another step that surely have to be computed is a 10-fold cross-validation to better assess the reliability of results.

However, an increase in accuracy can be registered when more covariates are involved in the model. Indeed, sensitivity in 2-covariates model results 83.33% more than 2 percentage points less than the 8-covariates model sensitivity. Furthermore, it is shown that in 8-covariates approach the probability of observing a changepoint in case patients is higher than the probability in same subjects in the 2-covariates model. These results can be appreciated comparing figures 5.18 and 5.19.

Future works include the development of involving covariates in the model that affect biomarkers trajectories directly on the baseline risk of disease. We are working on this method since we finished to draw up the method described in the current work. Baseline covariates are hypothesized to be factors that affect the probability of being a case. Prior prevalence $P(D_{N+1} = 1)$ is estimated from training data or another external source or a combination of both. Therefore a simply Bayesian logistic regression model can be conducted, independently from the model for the biomarkers. Therefore this method consists in implementing two separate models: one for longitudinal biomarkers trajectories and one for posterior risk of disease.

Last but not least, we just considered pre-clinical variables stable in time. A

very interesting development may be including time-varying covariates in the model, since they are supposed to have a crucial potential in order to increase world wide HCC survival rate.

We can conclude that overall covariates showed their potential in the current work. They could provide a critical increase in HCC early detection, and then in survival rate. The step to take now is finding the way to optimize their potential.

Bibliography

Nabihah Tayob, Francesco Stingo, Kim-Anh Do, Anna SF Lok, and Ziding Feng. A bayesian screening approach for hepatocellular carcinoma using multiple longitudinal biomarkers. *Biometrics*, 74(1):249–259, 2018.

Hashem B El-Serag and Andrew C Mason. Rising incidence of hepatocellular carcinoma in the united states. *New England Journal of Medicine*, 340(10): 745–750, 1999.

Hashem B. El-Serag. Hepatocellular carcinoma. *New England Journal of Medicine*, 365(12):1118–1127, 2011. doi: 10.1056/NEJMra1001683. URL `https://doi.org/10.1056/NEJMra1001683`. PMID: 21992124.

Jordi Bruix and Morris Sherman. Management of hepatocellular carcinoma. *Hepatology*, 42(5):1208–1236, 2005.

Steven J Skates, Donna K Pauler, and Ian J Jacobs. Screening based on the risk of cancer calculation from bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*, 96(454):429–439, 2001.

Martin W McIntosh and Nicole Urban. A parametric empirical bayes method for cancer screening using longitudinal observations of a biomarker. *Biostatistics*, 4(1):27–40, 2003.

Peter J Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

William MC Rosenberg, Michael Voelker, Robert Thiel, Michael Becka, Alastair Burt, Detlef Schuppan, Stefan Hubscher, Tania Roskams, Massimo

Pinzani, Michael JP Arthur, et al. Serum markers detect the presence of liver fibrosis: a cohort study. *Gastroenterology*, 127(6):1704–1713, 2004.

Wikipedia contributors. Alanine transaminase — Wikipedia, the free encyclopedia, 2019a. URL `https://en.wikipedia.org/w/index.php?title=Alanine transaminaseoldid=881303973`. [Online; accessed 23-June-2019].

Wikipedia contributors. Aspartate transaminase — Wikipedia, the free encyclopedia, 2019b. URL `https://en.wikipedia.org/w/index.php?title=Aspartate transaminaseoldid=884538098`. [Online; accessed 24-June-2019].

Margaret Sullivan Pepe, Ruth Etzioni, Ziding Feng, John D Potter, Mary Lou Thompson, Mark Thornquist, Marcy Winget, and Yutaka Yasui. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–1061, 2001.

Ringraziamenti

Ringrazio il mio relatore Francesco Claudio Stingo che si è reso disponibile a seguirmi sempre, anche dai posti più remoti del mondo.

Ringrazio l'Università di Firenze per avermi permesso di accedere alla macchina virtuale e "il super tecnico" Maria Nunzia Galdi che mi ha introdotto alla piattaforma di Google Cloud.

Ringrazio Matteo Pedone per avermi insegnato qualche trucco utile per parallelizzare.

Ringrazio Costi e i compagni di classe per avermi aiutato e reso i giorni in università molto meno pesanti; e Pallo che ha sempre cercato di risolvere ogni mio dubbio informatico e mi crede ora una Linux user.

Ringrazio Elena perchè mi ha fatto uscire allo scoperto.

Ringrazio Marti, Yako e gli amici di sempre per aver compreso il motivo delle mie assenze; i miei genitori e Bene per avermi sostenuto, soprattutto nel sofferto cambio di facoltà. E Marco, perchè c'è.