Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Longitudinal CEA measurements for prediction of colorectal cancer recurrence

Paul Mwangi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Roel BRAEKERS

SUPERVISOR :

Avishek CHATTERJEE

2020

2021

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Longitudinal CEA measurements for prediction of colorectal cancer recurrence

Paul Mwangi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Roel BRAEKERS

SUPERVISOR :

Avishek CHATTERJEE

Abstract

Background: Colorectal cancer (CRC) is the third most common type of cancer diagnosed worldwide, with about 1.4 million new cases each year. Surgery is the most frequently used treatment option, followed by a five-year post-surgery observation period to monitor the prognosis including the carcinoembryonic antigen (CEA) biomarkers. A rise in CEA levels may indicate a recurrence of the tumor.

Objectives: This study aimed to create a model to predict the CRC tumor recurrence using the baseline patient characteristics and longitudinal CEA measurement to help physicians make an optimal decision on individual medical care.

Methodology: A total of 2100 CRC patients who underwent surgery at Zyuderland Medical Centre (Netherlands) between 2008 and 2018 were followed up for five years. A joint model for time-to-event and longitudinal data was used. “Current value”, “current slope,” and “current value plus slope” parameterizations were used to link the time-to-event and longitudinal data. Time-dependent area under the curves (AUCs) and dynamic discrimination index (DDI) were used to evaluate the discrimination capability of the joint models for three and six-month intervals. The choice of best-fit model was based on the lowest Akaike information criterion (AIC) or Bayesian information criterion (BIC) and also a high DDI value was considered.

Results: 17.4% of the patients experienced tumor recurrence and 14.2% died after surgery. The results revealed that the “current value plus slope” parameterization had the highest discrimination power (DDI = 0.67) and lowest AIC (7849.92) and BIC (8007.28) values. The risk of tumor recurrence was significantly associated with current log CEA value (HR= 1.99, 95% CI: 1.66 - 2.41), the rate of change of the log CEA value (log hazard=12.29, 95% CI: 5.09 - 19.50), age above 75 years (HR= 1.41, 95% CI: 1.06 - 1.86), tumor stage three (HR= 1.99, 95% CI: 1.30 - 3.03), tumor stage four (HR= 7.91, 95% CI: 4.64 - 13.49), pre-surgery log CEA measurement (HR= 1.16, 95% CI: 1.03 - 1.32). Smoking status was not significantly associated with the risk of tumor recurrence.

Conclusion: In addition to their expertise, physicians can utilize a “current value plus slope” joint model formulation to help them make optimal medical care decisions.

Keywords: carcinoembryonic antigen, colorectal cancer, joint model, linear mixed model, cox proportional hazard model, time to event, area under the curve.

Acknowledgements

Firstly, I want to thank and praise the Almighty God for his blessings and for assisting me in completing my thesis successfully.

I am extremely grateful to VLIR-UOS for offering me a scholarship to study Master of Science in Statistics at Hasselt University in Belgium. I would not have been able to achieve my goals without their help.

I would like to express my sincere gratitude to my supervisors Prof Roel Braekers and Dr. Avishek Chatterjee. Words cannot express my appreciation for your acceptance to become my supervisor and for all the guidance, time invested, and the careful review of this successful work. I also want to thank the entire Center for Statistics (CENSTAT) fraternity for shaping me into the person I am today. Your generosity and diligence in sharing your knowledge and expertise throughout my entire course period were incredible.

A special thanks to Dr. Cary Oberije. Thank you for availing me of the opportunity to utilize data from the Zyuderland Medical Center in the Netherlands. This has not only demonstrated your faith in me, but it also provided another source of inspiration.

Finally, to all my fellow students and friends, your help and wonderful memories will be a part of my life story for the rest of my life.

I am speechless in gratitude for all my dear parents and siblings' prayers and the journey we have taken together thus far.

Give thanks to the lord, for he is good; his love endures forever

Contents

List of Tables

List of Figures

List of abbreviations and acronyms:

CRC: Colerectal Cancer

CEA: Carcinoembryonic Antigen

LMM: Linear Mixed Model

LDA: Longitudinal Data Analysis

JM: Joint Modelling

ROC: Receiver Operating Characteristic

AUC: Area Under the Curve

DDI: Dynamic Discrimination Index

HR: Hazard Rate

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

1 Introduction

1.1 Background

Colorectal cancer (CRC) is the third most common cancer diagnosed worldwide and one of the leading causes of cancer-related death, with about 1.4 million new cases and 700,000 deaths per year (Haggar *et al.*, 2009; Ferlay *et al.*, 2018). CRC is commonly found in the lining of the colon (large intestine) or the rectum, and it usually develops from focal changes within benign, precancerous polyps (Simon *et al.*, 2016). CRC results from the gradual accumulation of genetic and epigenetic changes that transform the normal colonic epithelium into cancer (Coppedè *et al.*, 2014). Age, sex (males), smoking status, excessive alcohol consumption, physical inactivity, high consumption of red and processed meat, obesity, and having a family history of CRC are the most commonly reported risk factors for CRC (Brenner *et al.*, 2018). In most cases, surgery has been the first line of treatment, with minimally invasive surgery becoming increasingly popular (Babaei *et al.*, 2016).

The burden of the CRC is a significant challenge in public health worldwide (Keum & Giovannucci, 2019). The incidence of CRC varies geographically, with the more-developed regions like Europe, Northern America, Australia, New Zealand, and Japan having a higher incidence compared to developing countries (Kuipers *et al.*, 2015). In Europe, there are about 3.91 million new cases of cancers and 1.93 million deaths, and CRC is among the most common type of cancer with 12.8% cases of the total (Kuipers *et al.*, 2015). Hungary (51.2), Slovakia (43.8), Norway (42.9), Slovenia (41.1), Denmark (41.0), Portugal (40.0), Netherlands (37.8), Belgium (35.3), Ireland (34.0), and Spain (33.4) were among the European countries with the highest age-standardised rates per 100 000 people in 2018 (Bray *et al.*, 2018).

Various types of research have been conducted to study the prevention of CRC tumor recurrence, and numerous lifestyle behavioral factors and biomarkers have been identified as important determinants. For example, physical activities and healthy dietary habits are among the primary protective factors of CRC recurrence (Brenner *et al.*, 2018; ACS, 2020). After surgical treatment of CRC, the patients are followed up for five years, where the standard follow-up includes the clinical examination, for example, to monitor the serum carcinoembryonic antigen (CEA) levels, imaging procedures such as ultrasound of the liver and computed tomography (CT) or PET-CT scan of the abdomen (Godhi *et al.*, 2017). Serum carcinoembryonic antigen (CEA) is a protein which normally is in low levels among adults, but its levels are often elevated in the presence of some cancer and non-cancerous (benign) conditions (CCS, 2021; Wang *et al.*, 2007). The CEA is the most commonly utilized biomarker test for CRC, and it is conducted as blood test to measure the amount of CEA protein in the blood. CRC develops slowly Brenner *et al.* (2013) and some studies have shown the continuous measurement of the CEA values can predict CRC tumor recurrence (Borges *et al.*, 2017; Peng *et al.*, 2015). Therefore, is of great importance to continue investing in statistical and epidemiological studies in CRC in order to study tumor recurrence.

In most cases, during the follow-up of the CRC patients, the time-to-event (e.g., tumor recurrence or death) data and the repeated measurements data (e.g., biomarker) among other clinical conditions are recorded for each patient. In the literature, methods for the separate analyses of the longitudinal outcomes and time-to-event are well documented, and these mainly include the Cox proportional hazard model for the survival outcome and mixed effect models for the repeated measurements (Cox, 1972; Molenberghs and Verbeke, 2005). However, the repeated measurements may be associated with the risk of an event of interest; hence modeling the survival and longitudinal data separately will not account for this association (Ibrahim *et al.*, 2010). An alternative solution is to use the repeated measurement as a time-varying covariate specified in the Cox regression model. Consequently, this increases the bias in parameter estimation, because the extended Cox regression model assumes the covariate can be measured all the time without error (Rizopoulos, 2012).

The joint modeling (JM) approach has been receiving more attention for the last years. Joint models for longitudinal and time-to-event data allow us to simultaneously model both the two processes to assess the association of repeated measurement and the time-to-event of the event of interest (Rizopoulos, 2012). JM has more advantages than the traditional methods because it provides efficient estimates of the covariates to the time-to-event and repeated measurements, and it also reduces the bias in the parameter estimation because it accounts for the measurement error (Ibrahim *et al.*, 2010). Also, JM allows for individual-specific predictions, which may help physicians make optimal decisions for individual patients (Lawrence *et al.*, 2015).

1.2 Rationale

Approximately 35% to 40% of CRC patients who receive surgery with or without chemotherapy, tumor recurrence may occur within 3 to 5 years of treatment (Colorectal Cancer Alliance, 2019; Guthrie, 2002). Usually, CRC patients are followed up with scheduled CEA testing for five years after surgery, and the clinical decisions to investigate further tumor recurrence using imaging are reached based on the CEA biomarker value and other clinical conditions (Shinkins *et al.*, 2017). An increase in CEA value may result in CRC tumor recurrence, but this is not always the case in some patients because CEA levels can increase despite the absence of recurrence, whereas in others, CEA values are high at the time of initial surgery but lower at the time of tumor recurrence (Saito *et al.*, 2016). As a result, the sensitivity and specificity of serum CEA for detecting recurrence are reported not to be high (Sorensen *et al.*, 2016). Despite its poor sensitivity and specificity in cases of early cancer, CEA can detect recurrence early in colorectal cancer with a continuous examination after surgery (Fletcher, 1996; Duffy *et al.*, 2013a). Also, the CEA is the most cost-effective way of detecting the recurrence of CRC in the primary care context (Mant *et al.*, 2013).

Identifying patients at high risk of tumor recurrence at an early stage would allow for more intensive follow-up on these patients, potentially allowing them to begin second-line treatment sooner. In order to contribute to the understanding of the progression of the CRC, we propose to develop a joint model for the longitudinal CEA measurement and tumor recurrence to study the CRC among patients from Zyuderland Medical Center, located in the Netherlands. We hypothesize that the tumor recurrence might depend on the current value of CEA plus the pace of slope trajectory at a specific time point, and therefore they would be incorporated in the model. Given that physicians utilize different types of information to predict patient prognosis, such as patient characteristics, medical history, and biomarkers such as CEA, the proposed prognostic model would not only be based on the longitudinal CEA measurements but also on other characteristics.

This project aims to refine and optimize the clinical decision for individual patients who received surgery using joint modeling. Particularly, we analyze the longitudinal CEA measures obtained in all blood tests for each patient diagnosed with CRC, tumor recurrence variable, and baseline characteristics (age, sex, tumor stage, tumor type, smoking and resection margin) from the Zyuderland Medical Center. As the follow-up continues, the clinicians take new CEA measurements, and therefore this accentuates that the joint model should be updated with these new measurements to aid in making appropriate intervention decisions.

The rest of the report is organized as follows. Section 1.3 describes the underlying research objective. Section 2 gives the details of the study setting, data description, and the main methodology employed in this report. Section 3 represents the results, Section 4 illustrates the discussion of the main results and finally, Section 5 represents conclusion and recommendation.

1.3 Research objective

To create a model to predict recurrence of the CRC tumor, using baseline patient characteristics and longitudinal CEA measurements.

2 Methods

2.1 Study setting and participants

This is a retrospective cohort study. Data used in this analysis were obtained from patients attending cancer treatment at Zyuderland Medical Center (in Netherlands) between 20th March 2003 and 8th October 2018. Specifically, these were patients who had surgery to remove colorectal cancer tumors. Zuyderland MC has specific expertise on oncology, obesity, neurocognitive diseases, mobility, and interstitial lung diseases. In collaboration with Maastricht University Medical Centre, among others, Zuyderland offers top clinical courses. Also, it serves as a regional teaching hospital for physicians. The Hospital has 980 beds, about 42000 submissions per year, 170000 nursing days per year, 86000 daycare submissions per year, and 845000 outpatient per year (Zuyderland, 2021).

The data set contained a total of 2301 patients. The CEA biomarker was measured for all patient in the follow up period after surgery at different time points. In addition, physicians performed liver echography or a CT scan or a PET-CT scan on occasion if the CEA levels rose quickly, or the patient was unwell, or other clinical signs. A rule of thumb was used for the decision to perform imaging. Patients with a CEA increase of 40% or 20% at two consecutive measurements were referred to a CT scan.

2.2 Data description

The data set contained the repeated CEA measurement, tumor recurrence, and death information. In addition, the following baseline characteristics variables were also available; age in years, smoking status, sex, tumor stage, tumor type, and resection margin (whether surgeon removed all tumor tissue). The analysis for this report only included those patients with CRC and had received their final surgery treatment between 1st January 2008 and 1st January 2018. The patients were excluded from the analysis if they corresponded to the following exclusion criteria; cancer stage 0, received palliative treatment, and received operation before 1st January 2008 and after 1st January 2018.

The measurements that were collected after tumor recurrence or death, or follow-up period not more than 5 years, were not included in the longitudinal data analysis because no participant's information was collected after death. Whereas in the case of tumor recurrence, the distribution of CEA measurement may change. Since the main purpose of this work was to predict CRC tumor recurrence, we treated death before tumor recurrence and all the observations collected after a five-year follow-up as a right-censored event-time. The time-to-event (in months) was the difference between the date of tumor recurrence from the final surgery date. A Pre-surgery CEA measurement was the value closest to the surgery date, and we concluded CEA value taken after more than six months before surgery as a missing value. The outcome variable CEA measurement overtime was log-transformed as follows; CEA$(Ug/L) = ln($CEA$(Ug/L)+1)$ to reduce the skewness of the data (Curran-Everett, 2018). We divided our data set into two parts; we randomly selected 80% of the total sample and named it "training data set," and the remaining 20% was named "test data set." The training data set was used to build joint model, whereas we used the test data set to assess the predictive ability of the joint model. Table 1 represents the variables used in the data analysis for this report.

Table 1: Covariates used in analyes of CEA data set

Variable	**Explanation**
Pre-surgery CEA	The closest CEA value before surgery
Follow-up	Date at measurement - Date at Final surgery
Tumor type	0 if adenocarcinoma, 1 if Mucinous carcinoma, and 2 if other
Tumor stage	1 if stage one, 2 if stage two, 3 if stage three, and 4 if stage four
Smoking	0 if never smoked, 1 if current smoker, and 2 if stopped smoking
Age in years	Date at final surgery - Date at birth
Sex	0 if male, 1 if female
Resection Merging	0 if no, if yes

2.3 Exploratory data analysis

In order to gain additional insights from the data, we conducted exploratory data analyses. The Kaplan-Meier survival analysis estimated the survival rate, and the log-rank test assessed the difference in the survival rate between the groups. Histograms examined the distribution of events over time. We used individual profile plots to investigate intra- and inter-patient variability for CEA measurement over time. Understanding data through data exploration facilitated the joint model building.

2.4 Joint model framework

The data used for this project contained both time-to-event data for tumor recurrence and longitudinal measures of tumor marker (CEA), and therefore joint modeling these two processes was deemed by us to be the best suited approach. The joint model is made up of two sub-models: a mixed effect sub-model for repeated measurement and a time to event sub-model for survival data, which are linked together via an association structure that assesses the relationship between the outcomes of interest (Ibrahim *et al.*, 2010; Wulfsohn and Tsiatis, 1997). The defining characteristic of a joint model is that survival and longitudinal data are modeled simultaneously with respect to a conditional density, instead of modeling them with two marginal and independent densities (Rizopoulos, 2012). The idea behind the joint model used for this study was that we used an appropriate random effect model to describe each patient's evolution of the CEA biomarker in time and then incorporated the estimated patient-specific evolution in the time-to-event model. In the next section, the longitudinal sub-model, the time-to-event sub-model, and parameterization options used for this study to associate the time-to-event (tumor recurrence) and the repeated CEA measurements in the joint modeling are discussed.

2.5 Joint model specification and formulation

The joint model formally associates the longitudinal and survival process through shared parameters (Henderson *et al.*, 2000; Rizopoulos, 2012). Therefore this model models the hazard of experiencing the event as dependent on the subject-specific characteristic of its longitudinal trajectory.

For the i^{th} subject, let T_i be the observed event time, T_i^* be the 'true' time-to-event, C_i is the censoring time, $T_i = min(T_i^*, C_i)$ be observed event time and δ_i is the event indicator where $\delta_i = 1$ if event; $\delta_i = 0$ if censored. Let $y_i(t)$ denote the longitudinal outcomes for subject i ($i = 1, \cdots, n$) taken at different time points time t ($t = 1, \cdots, T_i^*$). Longitudinal outcomes are often composed of observations of subjects that are measured repeatedly over time. For instance, the CEA values belonging to the same person were measured repeatedly at different time point. Based on the data exploration in Figure 5, the linear mixed model was the most reasonable model to consider. The longitudinal linear mixed effect submodel is in the form;

$$y_i(t) = m_i(t) + e_i(t) = x_i^T(t)\beta + z_i^T(t)b_i + e_i(t) \tag{1}$$

Where b_i is the subject-specific random effect which we assume a multivariate normally distribution, namely $b_i \sim N(0, D)$ (where D is a covariance matrix). $m_i(t)$ denotes the true and unobserved value for the longitudinal outcome at time t. $y_i(t)$ is the observed measured value which deviates from $m_i(t)$ by the amount of error $\epsilon_i(t)$, where $\epsilon_i(t) \sim N(0, \sigma^2)$. $x_i(t)$ and $z_i(t)$ depict the design matrix for the fixed effects β and random effects b_i (random intercepts and slopes), respectively. Random effects (random intercepts and slopes) express how the individual intercepts and slopes deviate from the average intercept and slope.

This report postulated different varying joint models for the association between the time to tumor recurrence and longitudinal CEA values. The survival submodel takes the form:

$$h_i(t|M_i(t), w_i) = h_0(t)exp[\beta^T w_i + f\{m_i(t), b_i, \alpha\}], \qquad t > 0, \tag{2}$$

where $M_i(t) = \{m_i(q), 0 \leq q < t\}$ represents the history of the true unobserved longitudinal process $m_i(t)$ up to time point t, and q is the time point prior to t. Parameter α quantifies the association of the underlying longitudinal process (e.g., CEA biomarker) at time t and the risk for an event (e.g., tumor recurrence) at the same time. $h_i(t)$ is the hazard for the i^{th} patient to experience the event of interest at time t. w_i is a vector of the baseline covariates with corresponding regression coefficients β. $h_0(t)$ is the baseline hazard when all covariates w_i are equal to zero. b_i is a vector of random effects for patient i. Various options of the function f are usually used to associate the time-to-the event submodel (Equation (2)) and the longitudinal submodel (Equation (1)) (Rizopoulos, 2012). This report highlights some of the options used to associate the two processes in the following paragraph.

Several options of function f include; "interaction effects", "lagged effects", "current true value plus the slope", "random effects" and "cumulative effect" parameterizations. "Interaction effects" parameterization assumes the current level of longitudinal measure is the same in all subgroups of the population under interest. "Lagged effect" parameterization assumes the risk of an event is associated with the repeated measurement at the previous time point $t - c$, where

c specifies the time lag of interest. "Cumulative effects" parameterization assumes the whole history of longitudinal measurement up to time t predicts the risk of experiencing an event at time t. "Random effects" parameterization only includes the random effect for the longitudinal submodel in the survival submodel (see Rizopoulos, 2012; Cekic *et al.*, 2019).

In particular, we used the "current true value" and "current true value plus the slope" parameterizations because it was hypothesized that the colorectal cancer tumor recurrence would depend on the current CEA value at time t and slope trajectory at the same time.

The "Current value" parameterization. This association assumes the risk for an event for the individual i at time t is associated with the true value $m_i(t)$ of the longitudinal outcome at the same time. However, this association does not differentiate between the individuals with an equal longitudinal score (e.g., biomarker) at a specific time point. The corresponding survival submodel has the form:

$$h_i(t|M_i(t), w_i) = h_0(t)exp\{\beta^T w_i + \alpha m_i(t)\}, \qquad t > 0, \tag{3}$$

Where α indicates the strength of association. Therefore the hazard of experiencing an event at time t depends on the true value of the longitudinal outcome at that time, baseline hazard, and baseline covariates.

The "Current value plus the slope" parameterization. This association extends the "current value" parameterization (Equation (3)) by adding the rate of change of the measurement at time t, which is estimated by the derivative of $m_i(t)$ with respect to time as shown in the Equation (5). The corresponding, relative risk sub-model has the form;

$$h_i(t|M_i(t), w_i) = h_0(t)exp\{\beta^T w_i + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}, \qquad t > 0, \tag{4}$$

Where

$$m_i^{'}(t) = \frac{d}{dt} m_i(t) = \frac{d}{dt}\{X_i^T(t)\beta + z_i^T(t) b_i\}, \tag{5}$$

This association structure assumes the hazard of experiencing an event at time t depends on both the current value of $m_i(t)$ and the slope of the current trajectory at time t ($m_i^{'}(t)$) (Ye *et al.*, 2008). The parameters α_1 is the association between the longitudinal current value (e.g., biomarker) with time-to-event at time t, and α_2 is the association between the value of the slope of the longitudinal outcome at time t with the time-to-event at the same time. This model can distinguish between the patients with the same biomarker value but with different slope trajectories at a specific time point. Also, the current slope trajectory association at a specific time point t can be obtained, excluding the current value association from Equation (4). The corresponding current slope association model is in the form;

$$h_i(t|M_i(t), w_i) = h_0(t)exp\{\beta^T w_i + \alpha_2 m_i'(t)\} \tag{6}$$

The Cox proportion model allows the baseline hazard function to be unspecified so that the misspecification of the distribution survival time is avoided but is not the same case for the joint model (Rizopoulos, 2012). Leaving the baseline hazard function unspecified in joint model can result in biased parameter estimates due to the underestimation of the standard errors (Rizopoulos, 2012; Lawrence *et al.*, 2015). The baseline hazard function can be estimated using non-parametric or parametric distributions (Kalbfleisch and Prentice, 2011). The most common parametric specification are the Weibull, the log-normal, and the gamma, while nonparametric specifications can be obtained using step functions and splines (Rizopoulos, 2012). Maximizing the likelihood is commonly used to estimate the parameters of the joint model.

2.6 Dynamic predictions

Frequently, physicians are interested in reliable prognostics about a patient to help them administer appropriate medical care. Therefore, based on the fitted joint model, conditional survival probabilities and projected longitudinal profiles are computed for a new subject i who has provided the set of longitudinal measurements up to a particular time point t. Let $y_i(t) = \{y_i(s), 0 \leq s \leq t\}$ be a set of longitudinal measurements for a new subject i, Rizopoulos (2011) considers based on the fitted joint model, conditional probability the of surviving up to time u (where $u > t$), given survival up to t, can be estimated as follows;

$$\pi_i(u|t) = pr\{T_i^* \geq u | T_i^* > t, y_i(t), w_i, D_n; \theta^*\} \tag{7}$$

Similarly, the predicted longitudinal outcomes at time u, is given by;

$$w_i(u|t) = E\{y_i(u)|T_i^* > y_i(t), D_n\} \tag{8}$$

Where $D_n = \{T_i, \delta_i, i = 1, \cdots, n\}$ represents the sample on which joint model was fitted, w_i represent the baseline covariates, θ^* represents the true parameter values and E denotes the expectation. The Bayesian approach is usually used in the computation of the probability because it eliminates the difficulties encountered when computing standard errors caused by the variability of both maximum likelihood and empirical Bayes estimates (Rizopoulos, 2011, 2012). The new information is added when new longitudinal measurements are recorded for subject i at time $t' > t$, where this information is used to update the predictions of $\pi_i(u|t')$ (conditional survival probabilities) and $w_i(u|t')$ (longitudinal outcomes), respectively (for details, see Rizopoulos, 2011, 2014).

2.7 Time dependent area under the curves (AUCs) and dynamic discrimination index (DDI)

It was of good interest to assess the predictive performance of the joint model in tumor recurrence. In this case, we were interested in the discriminative capability of the joint model within a given time window that was of medical relevance, in which it could distinguish those patients who would experience the tumor recurrence versus those who would not. Given the available longitudinal measurement $y_i(t)$ up to time t for subject j, the interest was to use this information on the medically relevant time frame $(t, t + \Delta t]$ in which a physician could make an appropriate intervention decision. Rizopoulos (2014), defined a prediction rule using $\pi_j(t + \Delta t|t)$, where, for c [0,1] (c denotes the threshold value at a specific time point), patient j is termed as the case (e.g., tumor recurrence) if $\pi_j(t + \Delta t|t) \leq c$ and as control (e.g., no tumor recurrence) if $\pi_j(t + \Delta t|t) > c$. Hence, the sensitivity and specificity is defined as;

$$\begin{aligned}\text{Sensitivity} &= P\{\pi_j(t + \Delta t|t) \leq c|T_j^* \in (t, t + \Delta t)\}, \qquad \text{and}\\ \text{Specificity} &= P\{\pi_j(t + \Delta t|t) > c|T_j^* > (t, t + \Delta t)\},\end{aligned}$$

The area under the curve (AUC) method based on Operating Characteristic (ROC) methodology is commonly used to assess the discriminative capability of models (Antolini *et al.*, 2005). The

AUC obtained for varying c is used to assess the discriminative capability of the joint model, and is given by;

$$AUC(t, \Delta t) = P[\pi_i(t + \Delta t|t) < \pi_j(t + \Delta t|t)]|\{T_i^* \in (t, t + \Delta t\} \cap \{T_j^* > (t, t + \Delta t]\}$$

Where i and j represent a pair of comparable patients who have provided the measurement (e.g, CEA values) up to time t. Here, the idea is that if we consider two patients, where one would experience the event (tumor recurrence), and the other will not experience the event at each time point and for a given time period. Then, joint model would assign a higher probability of not experiencing tumor recurrence beyond the selected time window for the patients who did not experience the tumor recurrence for a given time period of interest.

The AUC evaluates the discrimination accuracy of joint model at a particular time point. However, the dynamic discrimination index (DDI), which is the summary of AUCs, can be used to assess the overall discriminative capability of a biomarker at a given follow-up period (Njagi *et al.*, 2013). Rizopoulos (2014) proposed the following formula to compute the weighted average AUCs;

$$C_{dyn}^{\Delta t} = \frac{\int_0^\tau AUC(t, \Delta t) P\{\epsilon(t)\} dt}{\int_0^\tau P\{\epsilon(t)\} dt}$$

Where $\epsilon(t) = [\{T_i^* \in (t, \Delta t]\} \cap \{T_j^* > t + \Delta t\}]$, and $P\{\epsilon(t)\}$ denotes the probability of random pairs of subjects comparable at time t, and τ represent the main period of interest. $C_{dyn}^{\Delta t}$ is the dynamic discrimination index and it depends on the length Δt, which means that different models might have different discriminatory capability for different Δt. $AUC(t, \Delta t)$ estimate is based directly on its definition by properly counting the concordant pairs of subjects. The pair of subjects is concordant if the survival probability of subject i at time c is less than the survival probability of subject j. The estimation of $C_{dyn}^{\Delta t}$ is to obtain estimates for weight $P\{\epsilon(t)\}$ after $AUC(t, \Delta t)$ has been estimated. Note, the DDI do not fully account for censoring because the weighted proportion of pairs that cannot be compared due to censoring (Rizopoulos, 2011, 2014). A value of DDI value close to one implies the model has an excellent discrimination power.

2.8 The joint model set up for the CEA analysis

The data set used for this report comprised the baseline characteristics (age, sex, smoking status, tumor stage, tumor type and resection margin), and two outcomes, time-to-event (tumor recurrence), and the longitudinal CEA measurements. Before estimating the joint model, longitudinal and time-to-event submodels were selected independently. Hence, our analysis involved three main steps: (i) Separate longitudinal analysis of the repeated CEA measurement; (ii) Separate survival analysis of time to tumor recurrence using the Cox proportional hazard model; and (iii) Joint analysis of the repeated CEA measurement and time to tumor recurrence. Note, we used cross-validation to validate the joint model in this report. This method was adopted because it is one way to ensure the fitted joint model would be a good predictive model for tumor recurrence. Figure 1 shows the steps used to fit the joint model.

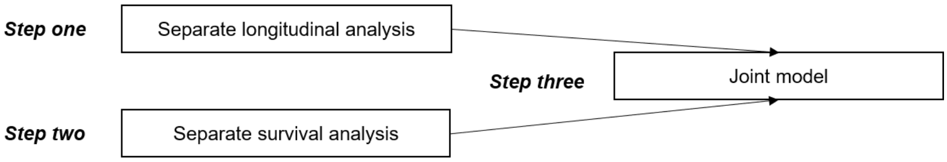


Figure 1: Steps used to fit joint model

In the *first step*: We built a linear mixed model with repeated CEA measurements as the outcome variable where the response was log transformed as follows (ln(CEA+1)). We incorporated random intercepts in the model since the individuals had a different CEA values after surgery. Smoking status and tumor stage covariates were added in the model as fixed effects because they significantly affect the CEA biomarker (Thota *et al.*, 2012; Beom *et al.*, 2020). Time in months was added as a random slope because individuals' CEA values were different over time. Besides, a mixture of chi-square was used to test whether the random slope was needed (Verbeke and Molenberghs, 2000). Also, we added sex and age into the model where likelihood ratio tests (LRT) were employed to assess their effect on the CEA levels. Note, in order to use LRT; we fitted the models using maximum likelihood estimation (ML) because the models with the same random effects (intercepts and slopes) and different fixed effects are comparable (Verbeke and Molenberghs, 2000). Maximum likelihood or restricted maximum likelihood (ML or REML, respectively) are typically used in the parameter estimation. In this report the REML was used in the parameter estimation because it produces the unbiased estimates of the variance and covariance parameters (Verbeke and Molenberghs, 2000).

In the *second step*, we analyzed the time to CRC tumor recurrence outcome using the Cox proportional hazard (PH) model. (Cox, 1972) proposed the Cox PH model, and this model is usually used to analyze the time-to-event data, and it assumes the baseline hazard functions of the covariates are proportional at a given time (Collett, 2015). The general form of cox model is:

$$h_i(t) = h_0(t) exp\{\beta^T w_i\}, \tag{9}$$

Where, $h_i(t)$ is the hazard for the i^{th} patient to experience the event of interest, e.g. tumor recurrence, at time t. $h_o(t)$ is the unspecified baseline hazard function at time t, given the reference category or 0 for all covariates. This baseline hazard function is assumed to be invariant across all the individuals meaning $h_o(t)$ does not depend on the individual i. β^T is the parameter effect which indicates how the hazard varies as a function of the explanatory variables w_i. In this analysis, age, sex, smoking status, tumor stage, tumor type, and resection margin were the potential prognostic factors for tumor recurrence. We included all these factors in the Cox PH model, where the outcome was the time to tumor recurrence. Also, we added the pre-surgery CEA measurement variable in the model, and the LRT was used to assess its effect. Moreover, the discriminative capability of the Cox PH model was evaluated using the concordance index (C index) proposed by (Harrell *et al.*, 1982). A model has an excellent discriminative power if it assigns patient i with a high-risk score than patient j given $T_i < T_j$. However, if pair of patients are censored, then C index does not consider them in computation. Values of C index close to 0.5 indicates poor prediction, and value close to one indicates excellent discrimination.

In the *third step*, the joint model consisting of Equations (1), and (2) was considered. First, baseline covariates were not included in the joint model to determine the effect of the longitudinal CEA biomarker on tumor recurrence. In this case, the longitudinal submodel only included the time in the fixed-effect structure, while in the survival submodel, it only incorporated the CEA biomarker's effect, as shown in Equation (13). Then, the baseline covariates were considered in the joint model. The "current value", "current value plus slope," and "current slope" parameterizations as discussed in the Equations (3), (4), and (6) were used to evaluate the association between the longitudinal CEA measurement and the time-to-tumor recurrence, respectively.

Moreover, a Weibull baseline hazard risk function $h_0(t) = \rho t^{\rho-1}$ was assumed. Where ρ is the shape parameter. Our primary motivation for using this parametric assumption was that it is usually more valuable and convenient when the objective is to obtain the absolute measure of the relative risk, such as predicting the outcomes for individual subjects (Lawrence *et al.*, 2015), which was in line with our objective.

2.9 Model diagnostics

The model diagnostic of the selected joint model was carried out by checking the diagnostics of the survival submodel and longitudinal submodel separately since the responses of these two parts are not comparable (Rizopoulos, 2012). For the Cox PH model (Equation (12)), Martingale residuals were calculated for the null model (a model with no explanatory variables) and plotted against the continuous covariates to determine the functional forms of these covariates, respectively. Next, we used the Schoenfeld Residuals test (SRT) to assess the PH assumptions of the Cox PH model. SRT determines whether the relationship between the scaled residuals and the time variable is significantly different from zero (Collett, 2015). Finally, we used the Cox-snell residuals to assess the overall fit of the joint survival submodel. In this case, the Kaplan-Meier plot of the Cox snell residuals estimates was compared with a unit exponential distribution survival function. If the assumed survival submodel fitted the data well, the cox snell residuals would follow a unit exponential distribution.

The assumptions of the linear mixed model are often assessed using subject-specific and marginal residuals (Verbeke and Molenberghs, 2000; Rizopoulos, 2010). However, the nonrandom nature of the dropout caused by the occurrence of events, affects these residuals for longitudinal processes. As a result, these residuals may not exhibit the standard properties of a linear mixed model, which can be misleading for validating joint model assumptions (Rizopoulos, 2010). Therefore, Rizopoulos (2010) proposed the multiple imputation residuals as a model assessment tool for linear mixed submodel. Further, under the complete data model, Rizopoulos (2010) proposed to supplement the observed data with randomly imputed longitudinal responses, corresponding to the longitudinal outcomes that would have been observed if the patients did not leave the study.

2.10 Model selection

The prediction of the joint models depends on the longitudinal biomarker itself in predicting the occurrence of the events and on the correct formulation of the joint models (Rizopoulos (2012)). Therefore, the Akaike information criterion (AIC) and Bayesian information criterion were used in conjunction with the dynamic discrimination index (DDI) to choose the best model. The AIC and BIC formulas are given by; AIC= $-2\ell(\hat{\theta}) + 2p$ and BIC=$-2\ell(\hat{\theta}) + p\ log(n)$ where p represents the number of parameters in the model, $\ell(\hat{\theta})$ is the log-likelihood of the fitted model and n is the number of observations. AIC tends to choose the most complex model while BIC tends to choose the most parsimonious model. The smaller values of BIC and AIC indicated the model had a better fit to the data .

2.11 Sensitivity analysis

It was important to assess the stability of the fitted joint models because the model's predictions are dependent on its specification and the biomarker itself (Rizopoulos, 2012). Therefore, we performed sensitivity analysis by re-specifying the baseline hazard with B-splines (see, Equation (A.18), in Appendix), with six knots spaced equally in the percentiles of the observed event times. B-splines are considered to be more flexible in most cases because increasing the number of knots may increase the flexibility in approximating the baseline hazard. In addition, we also linked the survival and longitudinal submodels using the lag of the current log CEA value, which assumed the risk of tumor recurrence was dependent on longitudinal log CEA at time t-c. This lagged effect survival submodel was expressed as follows:

$$h_i(t|M_i(t), w_i) = h_0(t)exp[\beta^T w_i + \alpha m_i\{max(t-c, 0)\}], \qquad t > 0, \tag{10}$$

In this case, we assumed the risk of tumor recurrence depended on previous time point earlier the current log value at time t. We used AIC and BIC to compare the various joint model formulations.

2.12 Handling the missing covariates

Missing data in the baseline covariates were assumed missing at random, meaning the missing values depended on the observed values but not on the unobserved values. Then, the missing data were imputed using the multiple imputation method because it accounts for the uncertainty given the imputed values were not observed (Groenwold *et al.*, 2012; Sterne *et al.*, 2009).

2.13 Software

The data management was done in Python and using open-source software for statistical computing and graphics, R version 4.0.5 (R Core Team, 2016). All the analyses were performed using R software. Under **nlme** package (Pinheiro *et al.*, 2017), a linear mixed model was fitted by **lme** function while **coxph** function under **survival** package (Therneau *et al.*, 2015) was employed to fit the survival submodel. **jointModel** function under **JM** package (Rizopoulos, 2010) was used to fit the joint model. 5% significance level was used for all data analyses.

3 Results

3.1 Exploratory data analysis

Figure 2 depicts the inclusion and exclusion criteria for the patients in this analysis. Of the total, 201 patients were excluded from the analysis, and of these, 30 patients had cancer stage 0, 80 patients received the palliative treatment, and 91 patients received surgery before 1st January 2008 or after 1st January 2018. In total, 2100 patients were included in the analysis because they corresponded to the following inclusion criteria; treated with surgery between 1st January 2008 and 1st January 2018.

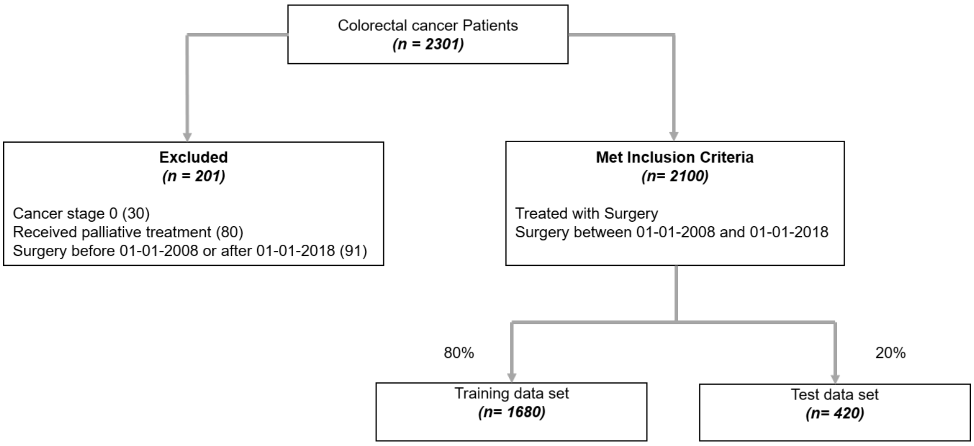

Colorectal cancer Patients
(n = 2301)
Excluded
(n = 201)
Cancer stage 0 (30)
Received palliative treatment (80)
Surgery before 01-01-2008 or after 01-01-2018 (91)
Met Inclusion Criteria
(n= 2100)
Treated with Surgery
Surgery between 01-01-2008 and 01-01-2018
80%
20%
Training data set
(n= 1680)
Test data set
(n= 420)

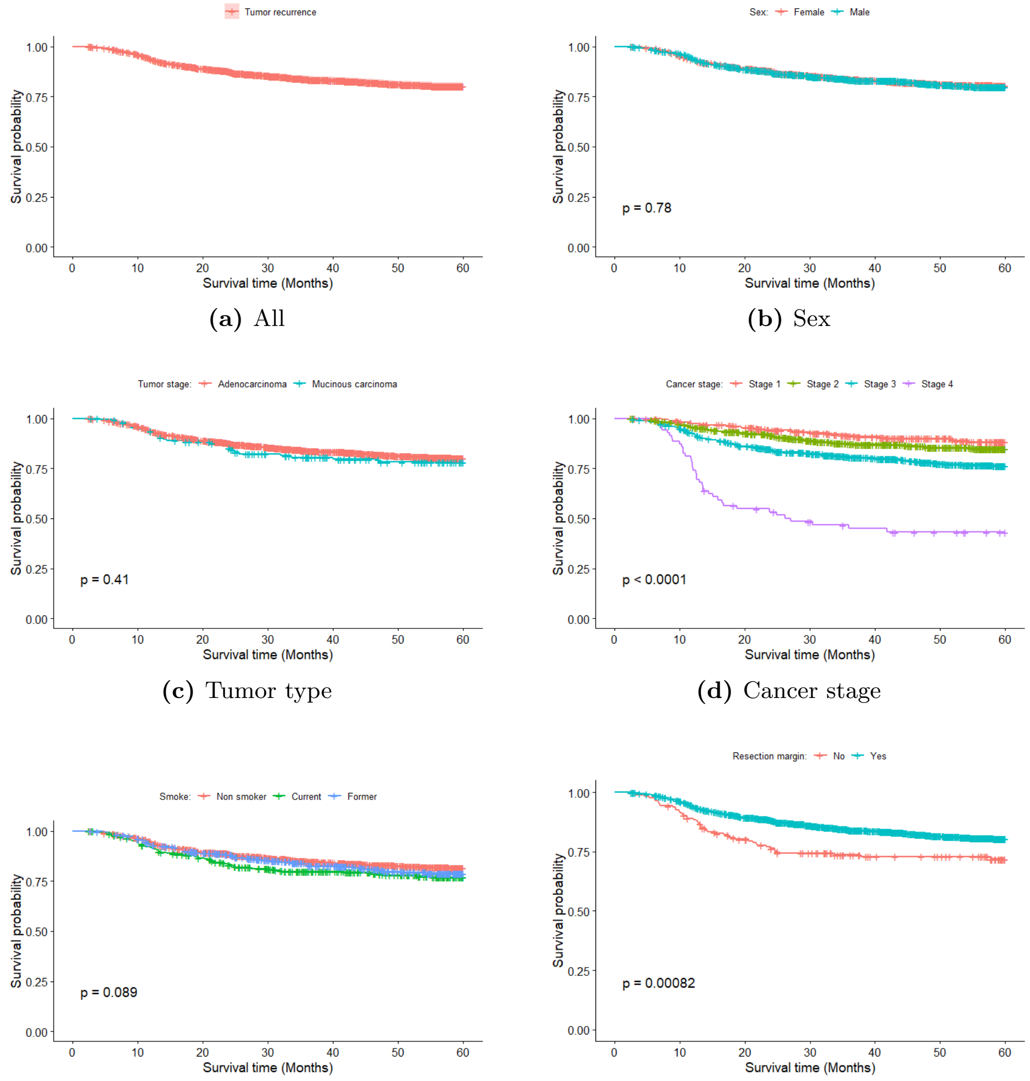
Figure 2: Inclusion and exclusion criteria flow chart

Table 2 summarizes sociodemographic characteristics (age, pre-surgery CEA measurement, tumor type, tumor stage, and resection margin) of study participants by tumor recurrence and death. Age at baseline was grouped as <= 75 or > 75 years old (<= 75 years as reference). For smoking status, patients were categorized as never smokers, former smokers (non smokers as reference). The resection margin was classified as yes or no (no as reference). The tumor stage was classified into stage one, stage two, stage three, and stage four (stage one as reference). Tumor type was grouped as adenocarcinoma and mucinous carcinoma (adenocarcinoma as reference). Of the total sample (n=2100), 366 (17.4%) of patients experienced the CRC tumor recurrence, and 299 (14.2%) patients died. The majority of the patients were females (56.2%), non-smokers (60.3%), had adenocarcinoma type of tumor (92.1%), were in cancer stage three (44.7%), and with resection margin (89.5%). Of the 2100 patients, 278, 112, 12, and 45 had their pre-surgery CEA measurement, smoking status, tumor type missing and resection margin, respectively. Their age at the first CEA measurement after surgery ranged from 62 to 75 years, with a median value equal to 69 years. The median follow up period was 18.9 months (IQR; 9.8 - 32.3). The number of visits per patient after surgery varied from 1 to 24, with a total of 19291 observations. The median of the number of visits was 10 (IQR; 6.0 - 9.2)

Table 2: Characteristics of participants by tumor, death and tumor or death (n=2100)

Baseline Characteristic	Overall n (%)	**Tumor recurrence**		**Death**		**Death or Tumor recurrence**	
		No	Yes	No	Yes	No	Yes
Total number of patients	2100	1734	366	1801	299	1614	486
Pre-surgery CEA (Median (IQR))	2.67 (1.47 - 5.10)	2.53 (1.40 - 4.73)	3.50 (1.99 - 8.08)	2.57 (1.40 - 4.93)	3.35 (2.10 - 6.81)	2.47 (1.40 - 4.60)	3.48 (1.99 - 7.66)
Age n (%)							
<= 75 years	1624 (7.3)	1346 (77.6)	278 (76)	1445 (80.2)	179 (59.9)	1285 (79.6)	339 (69.8)
> 75 years	476 (22.7)	388 (22.4)	88 (24)	356 (19.8)	120 (40.1)	329 (20.4)	147 (30.2)
Sex n (%), (*missing = 0*)							
Female	1183 (56.3)	974 (56.2)	209 (57.1)	1012 (56.2)	171 (57.2)	902 (55.9)	281 (57.8)
Male	917 (43.7)	760 (43.8)	157 (42.9)	789 (43.8)	128 (42.8)	712 (44.1)	205 (42.2)
Smoking n (%), (*missing = 112*)							
Current	336 (16.0)	269 (16.4)	67 (19.4)	278 (16.3)	58 (20.4)	244 (16.0)	92 (19.9)
Former	451 (21.5)	366 (22.3)	85 (24.6)	372 (21.8)	79 (27.7)	327 (21.4)	124 (26.8)
No	1201 (57.2)	1008 (61.3)	193 (55.9)	1053 (61.8)	148 (51.9)	955 (62.6)	256 (53.2)
Tumor type n (%), (*missing = 8*)							
Adenocarcinoma	1935 (92.1)	1600 (92.6)	335 (91.8)	1670 (93.0)	265 (89.2)	1495 (93.0)	440 (90.9)
Mucinous carcinoma	148 (7.1)	119 (6.9)	29 (7.9)	117 (6.5)	31 (10.5)	105 (6.5)	43 (8.9)
Other	9 (0.4)	8 (0.5)	1 (0.3)	8 (0.4)	1 (0.3)	8 (0.5)	1 (0.2)
Cancer stage n (%), (*missing = 0*)							
Stage one	372 (17.7)	337 (19.4)	35 (9.6)	340 (18.9)	32 (10.7)	316 (19.6)	56 (11.5)
Stage two	720 (34.3)	626 (36.1)	94 (25.7)	622 (34.5)	98 (32.8)	572 (35.4)	148 (30.5)
Stage three	939 (44.7)	740 (42.7)	199 (54.4)	793 (44.0)	146 (48.8)	696 (43.1)	243 (50.0)
Stage four	69 (3.3)	31 (1.8)	38 (10.4)	46 (2.6)	23 (7.7)	30 (1.9)	39 (8.0)
Resection margin n (%), (*missing = 45*)							
No	176 (8.4)	130 (7.7)	46 (12.6)	135 (7.7)	41 (13.9)	118 (7.5)	58 (12.1)
Yes	1879 (89.5)	1561 (92.3)	318 (87.4)	1625 (92.3)	254 (86.1)	1456 (92.5)	423 (87.9)

Figure 3 shows the overall and variable specific Kaplan-Meir survival curves of the CRC tumor recurrence, and each contained the *p*-value of the univariate log-rank test. The overall survival curve (Figure 3a) indicated a decrease in survival probability over time. The overall survival probability of the tumor recurrence was 93.0% at 1 year, 86.0% at 2 years, 83.5% at 3 years, 81.5% at 4 years and 80.0% at 5 years, respectively. There was no statistically significant difference seen between the survival curves of males and females (Figure (3b)). The patients with adenocarcinoma tumor type had a higher tumor free-recurrence survival than patients with mucinous carcinoma tumor type, although it was not statistically significant (Figure (3c)). Patients with cancer stage one had the highest free-recurrence survival probability, followed by cancer stage two, then cancer stage three, and patients with cancer stage four had the lowest free-recurrence survival probability, and this was statistically significant (Figure (3d)). In Figure (3e), the results revealed no significant difference in the free-recurrence survival probability among smokers, but the non-smokers had a higher probability of the disease-free response time. Those patients where the surgeon removed all their tumor tissue had a significantly higher free-recurrence survival probability than to those their tumor was not entirely removed during surgery (Figure (3f)). We observed a similar trend as previously discussed in the overall and variable specific Kaplan-Meir survival curves of the death event (Figure (A.1)) (in Appendix).



(a) All

(b) Sex

(c) Tumor type

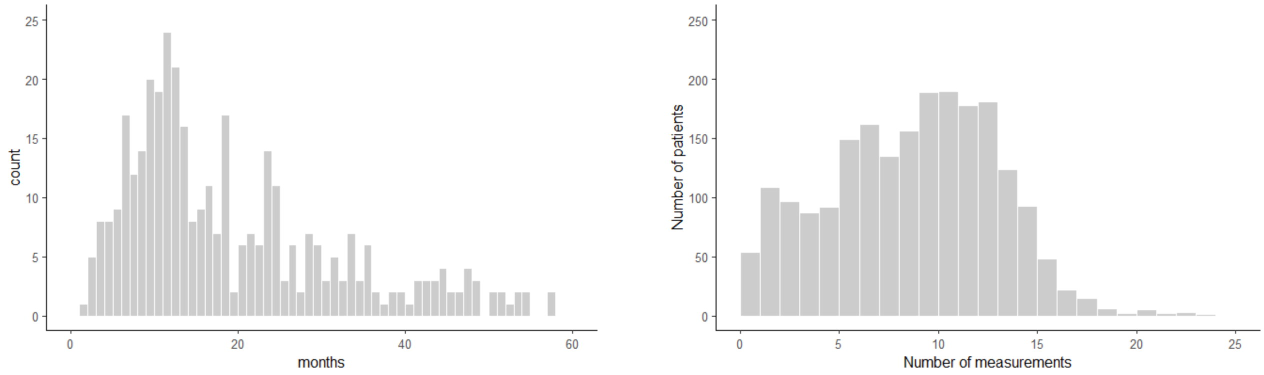
(d) Cancer stage

(e) Smoking status

(f) Resection margin

Figure 3: Overall and variable specific survival curves (*Tumor Recurrence*). For sex (b), Tumor type (c), Cancer stage (d), Smoking status (e), and resection margin (f). The p-value was obtained using the log-rank test

Figure (4a) depicts the histogram on the total number of patients per month who experienced tumor recurrence during the follow-up. It indicates that the majority of patients experienced tumor recurrence in the first 19 months after surgery. Figure (4b) shows a histogram of the number of patients at different number of CEA measurements. The number of CEA measurements after surgery varied from 1 to 24, and there were 19291 measurements of CEA values. The median number of CEA measurements was 10.

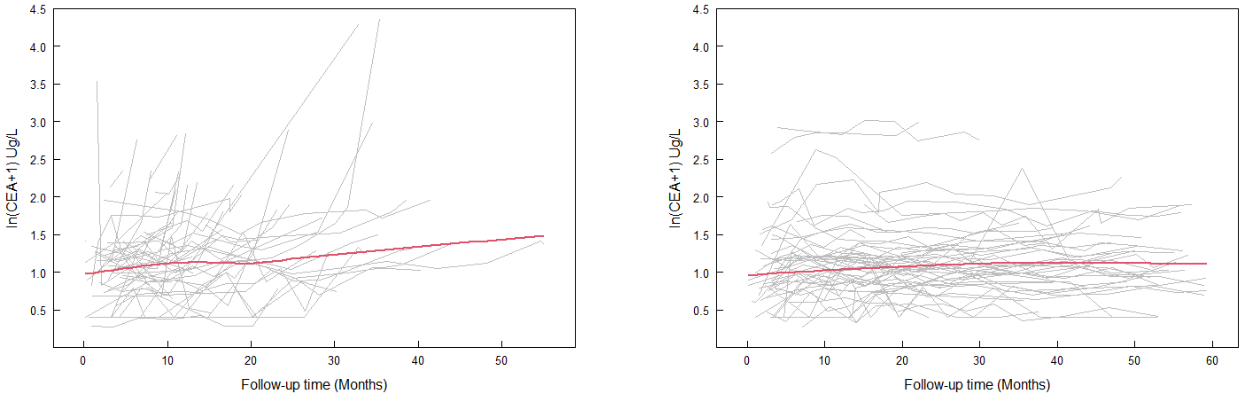


(a)

(b)

Figure 4: Figure (4a) is the histogram for the total number of patients per month who experienced tumor recurrence and Figure (4b) is the histogram for the total number of patients for different number of CEA measurement

Figure (5) represents the individual profile plots of the transformed CEA measurements against time in months. Figure (5a) shows the individual profiles for 50 randomly selected with tumor recurrence. It was observed that majority the patients had an increase of CEA value over time. Figure (5b) represents the individuals profiles for 50 randomly selected without tumor recurrence. It can be seen the CEA evolution for most of the individual profile almost remained constant during the follow up period. In general the plots indicated that for many individuals, their log CEA measurements seemed to be linear. Also, the individual profiles plots suggested that the CEA values started at different values after surgery, and there was variability within the subjects. Hence, a linear mixed model with random intercepts and slopes was the most plausible.



(a) Tumor recurrence

(b) No tumor recurrence

Figure 5: Figure (5a) represent the individual profiles for 50 randomly selected patients who experienced tumor recurrence. Figure (5b) represent the individual profiles for 50 randomly selected patients who didn't experienced tumor recurrence. The red line represents the loess smoother.

3.2 Linear mixed model results

First, the linear mixed model incorporated smoking status and tumor stage as fixed effects. Then, age and sex variables were added separately to the model, and the LRT results revealed that only age had a significant effect. Finally, using the LRT: $-2ln\lambda_N = -2(-3204.334{+}2209.152) = 1990.364 \sim \chi^2_{(0:1)}$, the random slope was found to affect the model significantly (p value <0.0001). Therefore, LMM with age, smoking status, tumor stage as the fixed covariate, and time as the random slope was considered. Note, we did not specify the correlation structure because the current JM R package works with a linear mixed model with no serial correlation structure. The package also assumes an unstructured covariance structure of the random effects (Rizopoulos, 2012). Thus, the final longitudinal submodel was in the form:

$$\begin{aligned} ln(y_i(t)+1) =& m_i(t) + e_i(t) \qquad (11) \\ =& \beta_0 + \beta_1 Age_i + \sum_{s=1}^{2} \beta_{2s} Smoke_s + \sum_{t=2}^{4} \beta_{3t} Tumor\_Stage_t \\ &+ \beta_4 Time + \beta_5 Time * Age_i + b_{i0} + b_{i1} Time + e_i(t) \end{aligned}$$

Here, $y_i(t)$ denotes the i^{th} patient CEA measurement at the t^{th} time point and this value is typically measured with the error. $m_i(t)$ represents the true unobserved CEA measurement at time t and $e_i(t)$ denotes the measurement error which is normally distributed with mean zero and variance σ^2. β_i denotes the effect of the baseline covariates. b_{i0} and b_{i1} represents the random intercepts and random slopes and they indicate the subject-specific deviations from the sample average intercept and average slope, respectively. Age$_i$= 1 for patients above 75 years and 0 for patients below or equal 75 years. Smoke$_s$ =0 for non smokers, 1 for current smokers and 2 for former smokers. Tumor stage$_t$= 1 for stage one, 2 for stage two, 3 for stage three and 4 for stage four. Type three F test Table A.1 (in Appendix) indicated that age, smoking status, tumor stage, time and interaction between age and time had a significant effect on the ln(CEA+1).

Table 3 depicts the linear mixed model results for the separate longitudinal analysis. The patients aged above than 75 years had a 8.3% higher expected ln(CEA+1) values than those aged below or equal 75 years old. For patients older than 75 years, the average ln(CEA+1) value increased by 0.59% for every one-month increase. The results revealed that the current and former smokers had 39.1% and 13.0% higher expected ln(CEA+1) values than non-smokers. The patients who had tumor stage four had 30.6% higher expected ln(CEA+1) values than patients who had tumor stage one. There were no statistical differences between the tumor stage two or tumor stage three and tumor stage one; respective p-values were 0.2881 and 0.0595.

Table 3: Paremeter estimates, 95% confidence interval (95% CI), standard error, and relative effects (RE) in separate linear mixed model fitted to the training dataset

Variable	**Estimates**	**95% CI**	**SE**	**p-value**	**RE**
(Intercept)	0.9262	(0.8638; 0.9886)	0.0318	<0.0001	2.5249
Age (above 75 years)	0.0801	(0.0197; 0.1405)	0.0308	0.0094	1.0834
Smoking					
Current smokers	0.3303	(0.2623; 0.3982)	0.0347	<0.0001	1.3913
Former smokers	0.1224	(0.0619; 0.1829)	0.0309	0.0001	1.1302
Tumor stage					
Two	0.0386	(-0.0325; 0.1097)	0.0363	0.2880	1.0393
Three	0.0659	(-0.0026; 0.1345)	0.0350	0.0596	1.0682
Four	0.2673	(0.1172; 0.4173)	0.0765	0.0005	1.3064
Time	0.0059	(0.0050; 0.0067)	0.0004	<0.0001	1.0059
Age*time	0.0025	(0.0006; 0.0044)	0.0010	0.0111	1.0025

3.3 Cox proportional hazard model results

In the Cox PH model, the potential prognostic factors were age, sex, smoking status, tumor stage, tumor type, and resection margin. Additional covariate such as pre-surgery CEA measurement was included in the Cox model, and the LRT results revealed it significantly impacted the model. Furthermore, using the martingale residuals plotted against the continuous pre-surgery CEA covariate revealed that linear assumption was not appropriate, and after log transforming, the linear assumption was met as shown in Figure A.3 (in Appendix). Also, using Schoenfeld's test revealed that all the variables met the proportional hazard assumption of the Cox PH model except resection margin Table A.2 (in Appendix). After stratification by resection margin, all variables met the PH assumption Table A.3 (in Appendix). Therefore, the final survival submodel was in the form:

$$\begin{aligned} h_i(t) = & h_0(t) exp\{\beta_1 Age_i + \beta_2 Sex_i + \sum_{s=1}^{2} \beta_{3s} Smoke_s + \beta_4 Tumor\_Type_i \\ & + \sum_{s=2}^{4} \beta_{5s} Tumor\_Stage_s + \beta_6 Pre\_Surgery\_log\_CEA\}, \end{aligned} \tag{12}$$

The results of the likelihood ratio test to assess the overall effect of the covariates in the Cox model revealed that age, tumor stage, pre-surgery log CEA measurement had a significant effect on the risk of tumor recurrence as shown in Table A.4 (in Appendix). Table 4 displays the results obtained from the Cox PH model for the tumor recurrence outcome. Patients aged above 75 years had a 39.0% higher risk of tumor recurrence compared to those age below or equal 75 years. For one unit increase in the pre-surgery log CEA, increased the hazard of tumor recurrence by 26.6%. Patients with tumors stage three and four, their risk of the tumor recurrence increased by 2.3 and 8.3 times, respectively compared to patients with tumor stage one. On the other hand, there were no differences between the current or former smokers and non-smokers, and the type of tumors; p-values were 0.0868, 0.4336, and 0.7192, respectively. Furthermore, the C index was calculated at 36 months and yielded a value of 0.6040.

Table 4: Parameter estimates, 95% confidence interval (95%), standard error (SE), hazard ratio (HR) in separate Cox PH model fitted to the training dataset

Variable	**Estimate**	**95% CI**	**SE**	**p-value**	**HR (95% CI)**
Age (above 75 years)	0.3294	(0.0533, 0.6055)	0.1409	0.0194	1.3901 (1.0548, 1.8322)
Sex (male)	-0.0464	(-0.2859, 0.1932)	0.1222	0.7045	0.9547 (0.7513, 1.2131)
Smoking					
Current smokers	0.2679	(-0.0387, 0.5744)	0.1564	0.0868	1.3072 (0.9620, 1.7761)
Former smokers	0.1157	(-0.1739, 0.4054)	0.1478	0.4336	1.1227 (0.8404, 1.4999)
Tumor type (Mucinous carcinoma)	-0.0859	(-0.5539, 0.3822)	0.2388	0.7192	0.9177 (0.5747, 1.4654)
Tumor stage					
Two	0.2546	(-0.1963, 0.7055)	0.2301	0.2685	1.2899 (0.8217, 2.0248)
Three	0.8185	(0.3977, 1.2392)	0.2146	0.0001	2.2670 (1.4885, 3.4527)
Four	2.1171	(1.5795, 2.6547)	0.2743	<0.0001	8.3069 (4.8523, 14.2210)
Pre-surgery log CEA	0.2362	(0.1257, 0.3467)	0.0564	<0.0001	1.2664 (1.1339, 1.4145)
C index at 36 months	0.6040				

C index: concordance index

3.4 Joint Models

The joint model was fitted by linking the estimated individual-specific evolutions' from the linear mixed model in the fitted Cox PH model as discussed in the methodology. We estimated the following joint models:

$$h_i(t) = \rho t^{\rho-1} exp\{\beta_0 + \alpha_1 m_i(t)\}, \tag{13}$$

$$\begin{aligned} h_i(t) = \rho t^{\rho-1} exp\{&\beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{s=1}^{2} \beta_{3s} Smoke_s + \beta_4 Tumor\_Type_i \\ &+ \sum_{s=2}^{4} \beta_{5s} Tumor\_Stage_s + \beta_6 Pre\_Surgery\_log\_CEA + \alpha_1 m_i(t)\}, \end{aligned} \tag{14}$$

$$\begin{aligned} h_i(t) = \rho t^{\rho-1} exp\{&\beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{s=1}^{2} \beta_{3s} Smoke_s + \beta_4 Tumor\_Type_i \\ &+ \sum_{s=2}^{4} \beta_{5s} Tumor\_Stage_s + \beta_6 Pre\_Surgery\_log\_CEA + \alpha_2 m_i'(t)\}, \end{aligned} \tag{15}$$

$$\begin{aligned} h_i(t) = \rho t^{\rho-1} exp\{&\beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{s=1}^{2} \beta_{3s} Smoke_s + \beta_4 Tumor\_Type_i \\ &+ \sum_{s=2}^{4} \beta_{5s} Tumor\_Stage_s + \beta_6 Pre\_Surgery\_log\_CEA + \alpha_1 m_i(t) + \alpha_2 m_i'(t)\}, \end{aligned} \tag{16}$$

$$\text{Where:} \quad m_i^{'}(t) = \frac{\partial(m_i(t))}{\partial(Time)} = \beta_5 Age_i + b_{i1} \tag{17}$$

The Weibull baseline risk hazard function $h_0(t) = \rho t^{\rho-1}$, was assumed, where ρ is the shape parameter and $exp(\beta_0)$ is the scale parameter. Equation (13) was used to evaluate the effect of the longitudinal log CEA measurement on the tumor recurrence by not considering the baseline covariates in both longitudinal and survival submodels. Equation (14) assumes that the risk for tumor recurrence at time t is associated with the current value of the log CEA measurement at the same time point. Equation (15) assumes the risk for tumor recurrence at time t is related to the slope of the current log CEA trajectory at the same time point. Equation (17) is the derivative of the longitudinal submodel (Equation (11)) with respect to time while Equation (16) postulates that the risk of the tumor recurrence depends on both the current value of the log CEA and the current slope of the log CEA trajectory. $h_i(t)$ is the hazard for the i^{th} patient to experience the the tumor recurrence, α_1 in Equation (13), (14) and (16) estimates the association between the current log CEA measurement at time t and the risk of tumor recurrence at the same time point and α_2 in Equation (15) and (16) quantifies the association of the current change of the log CEA measurements and the relative risk of the tumor recurrence.

3.4.1 Joint Model without baseline covariates

We first fitted the joint model without the baseline covariates as shown in Equation (13). The current value of the log CEA measurement was significantly predictive of the tumor recurrence (HR=2.6395, 95% CI: 2.4032 - 2.8990). This finding implied that the hazard of tumor recurrence was 2.6 times higher for every unit increase in the current log CEA value at a specific time point without considering the baseline covariates.

3.4.2 Association between the longitudinal and survival process

Different joint models formulation were considered; "current value", "current slope", "current value plus slope" parameterizations (Equation (14), (15), and (16)). The "current value" parameterization results revealed that the risk of tumor recurrence was 2.3 times higher for every one unit increase of the current log CEA value Table A.6 (in Appendix). Furthermore, in the "current slope" parameterization results, we observed that the slope trajectory of the log CEA was highly associated with the risk for tumor recurrence, and the corresponding log hazard ratio was 30.8016 (95% CI: 26.6870 - 34.9162) Table A.7 (in Appendix). For instance, this can translate that for every 0.02 unit change in the slope per month, the risk of tumor recurrence is associated with exp(30.8016×0.02) = 1.8515-fold (95% CI: 1.7052 - 2.0103) increase in the hazard. Also, the "current value plus slope" parameterization results revealed a significant association between the current value and the current rate of change of log CEA and the relative risk of the tumor recurrence. The hazard of the tumor recurrence increased by 99.8% for one unit increase in the current log CEA value while if the rate of the slope of log CEA measurement changed by 0.02 units, the risk of tumor recurrence increased by exp(12.2972×0.02) = 1.2788-fold (95% CI: 1.1073 - 1.4768) for patients having the same sex, smoking status, tumor type, and tumor stage (Table 6). Note, we did not use a one-unit change in the rate of change of current log CEA measurement slope trajectory because it is not meaningful, and it is enormous for the rate of change. From these three fitted joint models, we observed no material difference in their longitudinal submodel and survival submodel parameter estimates, respectively.

3.4.3 The time-dependent area under the curve (AUCs) and dynamic discrimination index (DDI)

We further evaluated how well the joint models would discriminate between patients who would experience the tumor recurrence and those who did not using the test data set. Finally, we computed the AUC$(t, \Delta t)$ for the 420 patients in the test data set. On many occasions, CRC patients visit the clinic for clinical examination every 3 to 6 months (Ryuk *et al.*, 2014; Godhi *et al.*, 2017). Therefore, AUCs were calculated at the follow-up times $\{t = 10, 20, 30, 40, 50, 60\}$ using $\Delta t = 3$ and $\Delta t = 6$, respectively. Also, most tumor recurrences may occur during the first two to three years after initial treatment (Jeffery *et al.*, 2016; Sargent *et al.*, 2005), and according to the Dutch guidelines, CEA testing is done every three to six months during the first three years and every six months for the remaining two years (Duineveld *et al.*, 2016). Therefore, in this analysis, the dynamic discrimination index (DDI) was computed for a follow-up period of 3 years or 36 months at a time length of 3 and 6 months, respectively. 200 number of Monte Carlo samples were used in the estimation of AUCs and DDIs, respectively.

Table A.9 (in Appendix) represents the results of the AUCs for the selected follow-ups times and the time considered to calculate AUCs at a certain time point. It can be observed that the AUCs are of varying degrees of the discriminative ability of the different joint models at different time points, where at time 50 months for $\Delta t = 3$ (Model III) had the highest AUC = 0.7956 and time 40 months for $\Delta t = 6$ (Model I) had the lowest AUC= 0.3044. Table 5 depict the results of dynamic discrimination index (DDI) for first 3 years of follow-up, AIC and BIC for the fitted joint models. The DDI results revealed that using the time window of 3 months had a better prediction of the patient who would experience the tumor recurrence versus not than a time window of 6 months in all four fitted joint models, respectively.

In contrast, to the C index computed at 36 months for the Cox model, all joint model formulations had a better discrimination power using the DDI considering the same follow-up period.

Table 5: Internal validity results showing the dynamic discriminative index (DDI) computed using the test dataset. Bayesian information criterion (BIC) and Akaike information criterion (AIC) values for four joint models fitted to the training dataset.

Joint model parameterization	**Time window t (Δt)**	**DDI (0-36 months)**	**AIC**	**BIC**
I: Current value	3	0.6363	10049.260	10087.240
(No baseline covariates)	6	0.6332		
II: Current value	3	0.6671	7868.916	8020.859
	6	0.6531		
III: current slope	3	0.6674	7890.566	8042.509
	6	0.6422		
IV: Current value plus slope	3	0.6706	7849.915	8007.284
	6	0.6553		

Model I: no baseline covariates in both survival and longitudinal submodels

Model II, III, IV includes the baseline covariates in both survival and longitudinal submodels

3.4.4 Joint Model selection

The AIC and BIC alongside with DDI were consinder to settle on the best joint model formulation. Based on AIC, BIC and DDI results as shown in Table 5, there was evidence that the joint model (IV) which includes the current log CEA value plus log CEA slope trajectory association as predictors of the risk of tumor recurrence was the best fitting joint model and higher discrimination power compared to joint model without baseline covariates, "current value" and "current slope" parameterizations, respectively. Therefore, we considered the "current value plus slope" joint model formulation as our best model, and was used to interpret final results and illustration of the dynamic predictions.

3.4.5 Joint model diagnostics

We further assessed the diagnostics of the best fitting joint model. Finally, we used the Cox Snell residuals plot to evaluate the overall fit of the survival submodel. Figure 6 shows the Kaplan-Meier estimates of the Cox Snell residuals. It can be seen that the gray line does not entirely hover through the solid line, especially for the residuals below 1.5. This plot suggests the survival submodel did not fully fit the data. Although we can argue the model did fit the data well because the unit exponential distribution lies within the 95% pointwise confidence intervals apart from residuals below 0.5. However, we did not evaluate the assumption linear mixed submodel because the residuals plots will not be reliable due to the non-random dropout of patients caused by tumor recurrence or death.

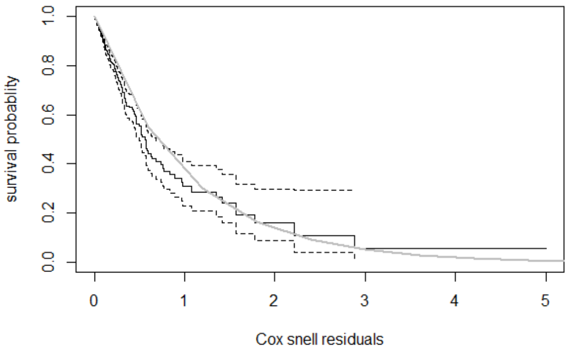


Figure 6: Cox Snell residuals plot. The black solid line denote the Kaplan Meier estimates of the survival functions of the Cox-Snell residuals. Dashed lines corresponds to the 95% pointwise confidence intervals. The gray line represent the survival function of the unit exponential distribution

3.4.6 Results for the "current value plus slope" parameterization

Table 6 represents the parameter estimates and the 95% confidence interval results for the "current value plus slope" parameterization. The results of the longitudinal sub-model were consistent with the results of the separate longitudinal analysis, as shown in Table 3. Thus, there was a slight difference, but there was no difference in their statistical significance. This similarity is because we used maximum likelihood to estimate the parameters in both models. Therefore, the interpretation of the results remains the same.

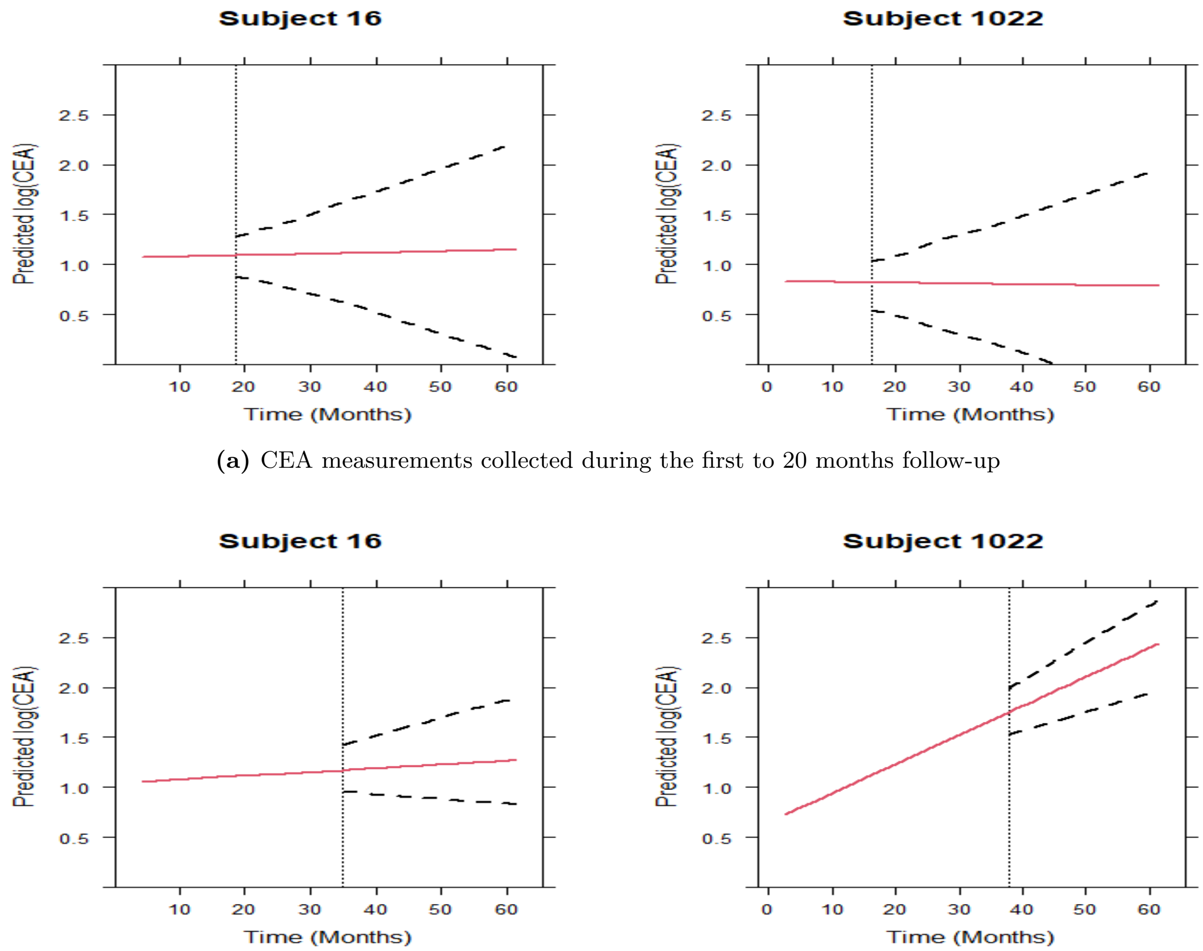
In contrast, in the survival submodel, the results were different from the separate survival analysis. This difference is because the joint model takes into account the measurement error in the CEA measurements. The likelihood ratio test, which was used to assess the overall effect of the covariate in the survival submodel, revealed that age, tumor stage, pre-surgery log CEA values, current log CEA value, and the rate of change of log CEA all significantly affected the risk of tumor recurrence as shown in Table A.8 (in Appendix). Males had a 33.1% lower risk of tumor recurrence compared to females. For patients who had tumor stage four during surgery, their hazard of the tumor recurrence increased by 7.9 times, while patients who had tumor stage three, their risk of the tumor increased by 2.0 times compared to those who had tumor stage one, respectively. The risk of tumor recurrence increased by 92.9% for a unit increase in the log CEA value before the surgery. There was no significant risk difference between current or former smokers and non-smokers.

Table 6: Parameter estimates and 95% confidence intervals for the "current value plus slope" parameterization results fitted to the training dataset

Variable	**Estimate**	**95% CI**	**SE**	**p-value**	**RE**
Longitudinal Sub-model					
(Intercept)	0.9229	(0.8606; 0.9853)	0.0318	<0.0001	2.5166
Age (above 75 years)	0.0789	(0.0190; 0.1389)	0.0306	0.0099	1.0821
Smoking					
Current smokers	0.3308	(0.2628; 0.3988)	0.0347	<0.0001	1.3921
Former smokers	0.1247	(0.0642; 0.1851)	0.0308	0.0001	1.1328
Tumor stage					
Two	0.0386	(-0.0324; 0.1097)	0.0363	0.2866	1.0394
Three	0.0646	(-0.0039; 0.1332)	0.0350	0.0647	1.0667
Four	0.2586	(0.1077; 0.4095)	0.0770	0.0008	1.2951
Time	0.0064	(0.0055; 0.0073)	0.0005	<0.0001	1.0064
Age:Time	0.0026	(0.0007; 0.0046)	0.0010	0.0086	1.0026
Variable	**Estimate**	**95% CI**	**SE**	**p-value**	**HR (95% CI)**
Survival Sub-model					
Intercept	-7.1357	(-7.7514; -6.5199)	0.3142	<0.0001	-
Age (above 75 years)	0.3410	(0.0619; 0.6200)	0.1424	0.0166	1.4063 (1.0638; 1.8590)
Sex (male)	-0.2660	(-0.5207; -0.0114)	0.1299	0.0406	0.7664 (0.5941; 0.9886)
Smoking					
Current smokers	0.2182	(-0.1013; 0.5378)	0.1630	0.1806	1.2439 (0.9037; 1.7122)
Former smokers	0.2099	(-0.0880; 0.5078)	0.1520	0.1673	1.2335 (0.9157; 1.6616)
Tumor type (Mucinous carcinoma)	-0.0089	(-0.4801; 0.4623)	0.2404	0.9704	0.9911 (0.6187; 1.5877)
Tumor stage					
Two	0.2412	(-0.2096; 0.6919)	0.2300	0.2943	1.2727 (0.8109; 1.9975)
Three	0.6868	(0.2638; 1.1098)	0.2158	0.0015	1.9874 (1.3019; 3.0337)
Four	2.0687	(1.5355; 2.6018)	0.2720	<0.0001	7.9143 (4.6438; 13.4882)
Pre-surgery log CEA	0.1507	(0.0263; 0.2752)	0.0635	0.0176	1.1627 (1.0266; 1.3168)
Current value	0.6919	(0.5049; 0.8789)	0.0954	<0.0001	1.9975 (1.6567; 2.4084)
Slope association	12.2972	(5.0972; 19.4971)	3.6734	0.0008	

3.4.7 Dynamic predictions

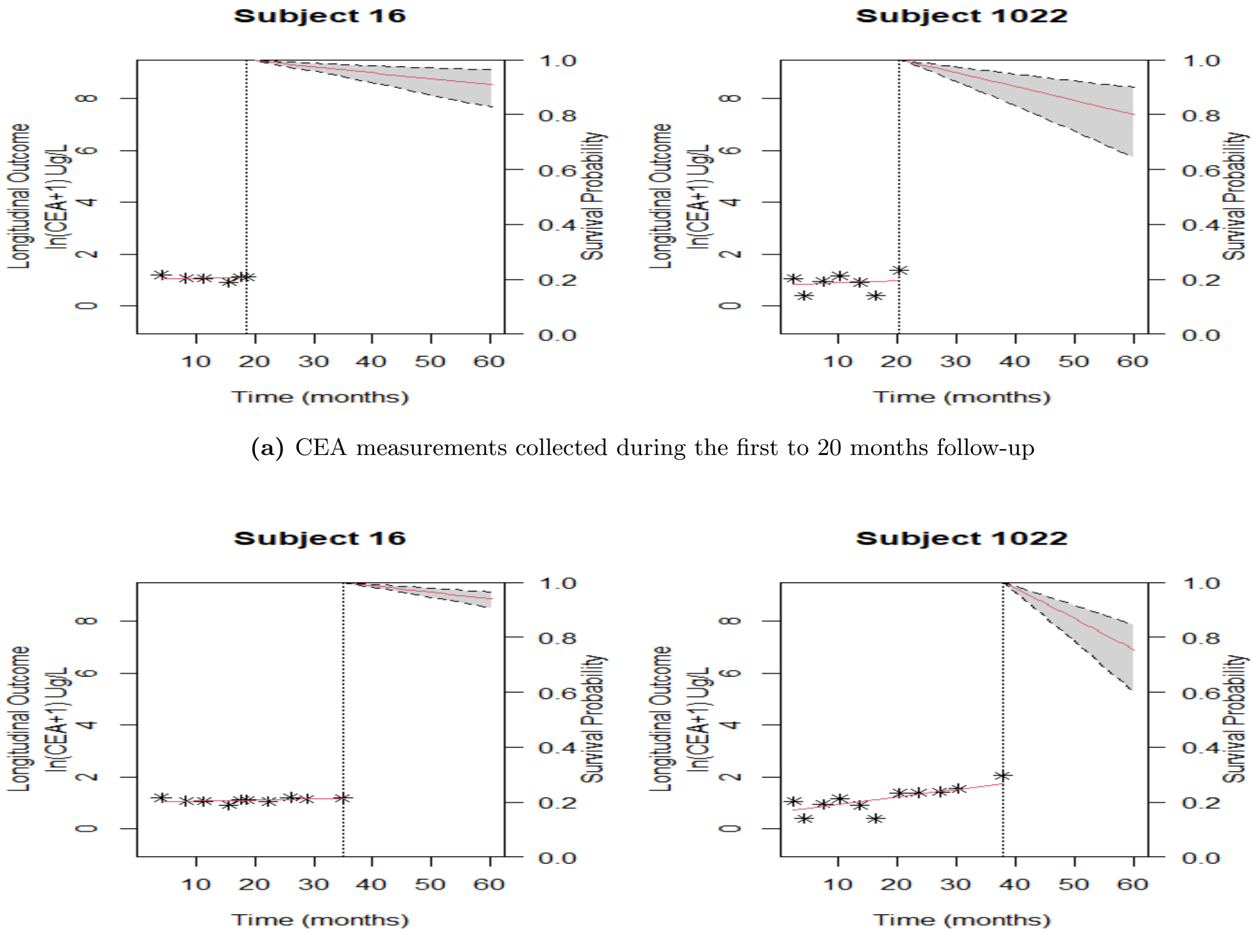
We considered the dynamic predictions for two randomly selected patients (subject 16 and subject 1022) from the test data based on the best fitted joint model ("current value plus slope" parameterization). These two patients had different baseline characteristics and longitudinal CEA measurements. Also, these patients had more than 10 visits, and they had provided the longitudinal CEA measurement for the first 40 months. Therefore, we considered the first 20 and 40 months for each patient because they had survived until this time. The conditional survival probability and the log CEA measurement trajectory predictions were then calculated for each patient for the remaining time up to the end of the follow-up period. Figure 7 represents the prediction of the log measurement trajectory for the two subjects using different time points until end of follow-up period, respectively. In general, subject 1022 had slightly higher increasing log CEA measurements, and the prediction of his log CEA measurement trajectory was higher than subject 16. Figure 8 shows the conditional probability plots at each of the remaining time points until end of follow-up period. Patient 16 had a more stable log CEA measurement profile; hence, he had a higher recurrence-free survival probability than patients 1022, who had a slightly increasing log CEA measurement trajectory. This finding seemed logical because the continuous increase of the CEA measurement may indicate tumor recurrence. These prediction plots illustrated how the joint model could help physicians make subject-specific decisions in terms of medical care.



(a) CEA measurements collected during the first to 20 months follow-up

(b) CEA measurements collected during the first to 40 months follow-up

Figure 7: Predicted longitudinal CEA trajectory (with a 95% pointwise confidence interval) for Patient 16 and 1022 from the test dataset. The dotted line denotes the last time point Patient 16 and 1022 were still event-free, respectively. The dashed line represent the 95% pointwise confidence interval.



(a) CEA measurements collected during the first to 20 months follow-up

(b) CEA measurements collected during the first to 40 months follow-up

Figure 8: Estimated conditional survival probabilities for Patient 16 and 1022 from the test dataset. The vertical dotted line represents the time point of the last log CEA measurement. The stars indicate the observed longitudinal data. The red line on the left of dotted line indicates the fitted log CEA measurement while the red line on the right represents the conditional survival probability and the dashed lines is the corresponding 95% pointwise confidence intervals

3.4.8 Sensitivity analysis

Finally, we tested the stability of the "current value", "current slope", and "current value plus slope" joint model formulations shown in Equation (14), (15), and (16), but with a B-splines baseline hazard function. Furthermore, we assumed that the risk of tumor recurrence was dependent on the current log CEA value one month before (Equation (A.19), in Appendix); thus, we fitted the model with the Weibull baseline hazard function and the B-splines baseline hazard function, respectively. Table 7 shows the AIC, BIC, and DDI for different joint model formulations. The results revealed that all of the fitted models had a DDI greater than 0.6. Thus, according to the results, AIC chooses the "current value plus slope" joint model formulation with the Weibull baseline hazard function (IV). In contrast, BIC selects the lagged effect joint model formulation (V) with Weibull baseline hazard function. Based on AIC and DDI, the Weibull baseline hazard joint model (IV), which incorporated both current log CEA values and slopes trajectory as predictors of the risk of tumor recurrence, had the best data fit and discrimination power.

Table 7: Dynamic discriminative index (DDI) computed using the test dataset. Bayesian information criterion (BIC) and Akaike information criterion (AIC) values for four joint models fitted to the training dataset.

Joint model parameterization	**Baseline hazard**	**Time window Δt**	**DDI (0-36 months)**	**AIC**	**BIC**
II: Current value	Weibull	3	0.6671	7868.916	8020.859
		6	0.6531		
	B-splines	3	0.6488	9555.423	9767.059
		6	0.6016		
III: Current slope	Weibull	3	0.6674	7890.566	8042.509
		6	0.6422		
	B-splines	3	0.6497	9501.205	9712.84
		6	0.6038		
Iv: Current value plus slope	Weibull	3	0.6706	7849.915	8007.284
		6	0.6553		
	B-splines	3	0.6473	9528.629	9745.691
		6	0.6021		
V: Lagged effect	Weibull	3	0.6664	7859.299	7994.963
		6	0.6536		
	B-splines	3	0.6487	9558.346	9769.981
		6	0.6001		

4 Discussion

Studies utilizing the joint models in predicting the CRC tumor recurrence are limited, and to the best of our knowledge, we found none in the literature. This report aimed to create a joint model for time-to-tumor recurrence and longitudinal CEA measurements while considering the baseline covariates to help the physicians make the optimal decision on individual medical care. The results demonstrated the usefulness of the repeated CEA measurement in the prediction of tumor recurrence. We found that baseline covariates age, sex, tumor stage, and the pre-surgery CEA measurement baseline covariates significantly affected the risk of tumor recurrence. Also, the results revealed that the risk of the tumor recurrence depended on the current log CEA value, current log CEA slope trajectory, and current log CEA value plus the current log CEA slope trajectory. Furthermore, the dynamic predictions plots demonstrated how a new CEA measurement could be utilized in the joint model to help physicians make the right decisions on the future medical care for individual patients. Using the baseline covariates and the longitudinal CEA measurements showed a good discriminative capability, with the DDI value for the three-year follow-up ranged above 0.6.

One of the key findings of this study is that it confirms repeated CEA measurements after CRC surgery is a good predictor of cancer recurrence after adjusting for the baseline covariates. Past studies also reported similar findings using different approaches of analysis (Kwaan *et al.*, 2020). In our findings, only 17.4% of patients experienced tumor recurrence after surgery. A study by Wieldraaijer *et al.* (2018) which looked at the clinical pattern of recurrence during the follow-up in the Netherlands, found a similar percentage of patients who experienced tumor recurrence. Also, similar findings have been reported by Primeau (2018). In this study, the cancer stage influenced the tumor recurrence, consistent with the results found by Azzam *et al.* (2020). Also, patients aged above 75 years had a high risk of tumor recurrence, which was consistent with the results reported by Macrae (2016). Our study found the males had a lower risk of tumor recurrence than females, contrary to what has been reported by previous studies (Ferlay *et al.*, 2015; Brenner *et al.*, 2018). On the other hand, some studies have shown that older women above 65 have a lower recurrence-free survival probability than men (Park *et al.*, 2013; Hansen *et al.*, 2012; Benedix *et al.*, 2010). In our study, most patients (63.2%) were aged above 65 years, which might explain why males had a lower risk in our finding than females.

However, we cannot justify our finding of males having a lower risk, and it needs further research.

Another key finding this study is that the joint model estimates were more robust than the estimates of the Cox PH model. Intuitively, this finding might be as result of the joint model producing less biased estimates. Corroboratively, a study by Powney *et al.* (2014) reported that the use of the joint model showed better significant difference in predicting treatment outcomes than when prediction was done in the separate analysis. Indeed, joint models have been applied in the literature to investigates the link between longitudinal biomarkers and events of interest on numerous diseases like cancer, HIV/AIDS, transplant studies, among others (Proust *et al.*, 2009; Brombin *et al.*, 2016; Abdi *et al.*, 2013). The preference for joint models partly stems from advantages such as reduction of bias in parameter estimation, account for the intermittent missing data in repeated measurements, and the inclusion of longitudinal covariates measured with errors into the survival submodel (Ibrahim *et al.*, 2010).

This study's main strength was sufficient sample size availability and the time-to-tumor recurrence, and the longitudinal CEA measurement was analyzed simultaneously while considering a breadth of risk factors. Hence, this led to an increase of power in predicting tumor recurrence and reduction of bias in estimating the repeated CEA measurement effect on the CRC recurrence. However, the proposed methodology had several limitations. Only tumor recurrence-free survival was considered, and from the clinical point of view, the doctors might be interested in disease-free progression and overall survival. Besides, not all potential risk factors like lifestyle, medical history, genetic data, physical activities, and diet, as reported by Liang *et al.* (2020) and Primeau (2018), were available for analysis, and that might influence the decision making. The discrimination accuracy of the joint models was assessed via internal validation, which has some drawbacks as the model cannot be used outside the study setting. In this report, we assumed linear trajectories for the longitudinal CEA measurement, but this might not be the case for all individuals. However, the splines, nonlinear mixed models among others, can be considered if patients have highly nonlinear evolution (Desmée *et al.*, 2017).

5 Conclusion and recommendation

Our findings show that the “current value plus slope” joint model formulation was more reliable and offered better discrimination power than other joint models’ formulations. It appeared that this joint model was a valuable model for predicting the risk of tumor recurrence, and physicians can apply it to help in deciding on personalized medical care. However, this model should be used with some caution and should not replace physician expertise because other special individual characteristics might influence the risk of tumor recurrence. Besides, it has been shown that combining different biomarkers like carbohydrate antigen (CA) 19-9, CA72-4, and CA125 can have a higher sensitivity and specificity than using the CEA alone (Gao *et al.*, 2018; Wu *et al.*, 2020). Therefore, we recommend that future research studies should incorporate more biomarkers in multivariate joint modeling to enhance the accuracy of the predictions. To improve the utility of such models, for medical practice and among other professionals, we recommend that it is necessary to integrate the joint model in an interactive web-based platform where such professionals are able to update the patient’s information and predict health outcomes in a more simplified and user-friendly manner that is beneficial for decision making on individual medical care.

References

Abdi, Z. D., Essig, M., Rizopoulos, D., Le Meur, Y., Prémaud, A., Woillard, J. B., ... Rousseau, A. (2013). Impact of longitudinal exposure to mycophenolic acid on acute rejection in renal-transplant recipients using a joint modeling approach. *Pharmacological Research*, 72, 52-60.

Antolini, L., Boracchi, P., Biganzoli, E.. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24), 3927–3944. https://doi.org/10.1002/sim.2427

American Cancer Society(ACS) (2020). Colorectal Cancer Facts Figures 2020-2022. Atlanta: American Cancer Society;

Azzam, N., Alruthia, Y., Alharbi, O., Aljebreen, A., Almadi, M., Alarfaj, M., Alsaleh, K., Almasoud, A., Alsharidah, M., Alseneidi, S., Alali, F., Alalwan, M.. (2020). Predictors of Survival Among Colorectal Cancer Patients in a Low Incidence Area. *Cancer Management and Research*, Volume 12, 451–459. https://doi.org/10.2147/cmar.s233215

Babaei, M., Balavarca, Y., Jansen, L., Gondos, A., Lemmens, V., Sjövall, A. (2016). Minimally invasive colorectal cancer surgery in Europe: implementation and outcomes. *Medicine*, 95(22).

Benedix, F., Kube, R., Meyer, F., Schmidt, U., Gastinger, I., Lippert, H., Colon/Rectum Carcinomas (Primary Tumor) Study Group. (2010). Comparison of 17,641 patients with right- and left-sided colon cancer: differences in epidemiology, perioperative course, histology, and survival. *Diseases of the Colon Rectum, 53*(1), 57-64.

Beom, S. H., Shin, S. J., Kim, C. G., Kim, J. H., Hur, H., Min, B. S., ... Ahn, J. B. (2020). Clinical Significance of Preoperative Serum Carcinoembryonic Antigen Within the Normal Range in Colorectal Cancer Patients Undergoing Curative Resection. *Annals of surgical Oncology*, 1-10.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: *A Cancer Journal for Clinicians*, 68(6), 394-424.

Brenner, H., Chen, C. (2018). The colorectal cancer epidemic: challenges and opportunities for primary, secondary and tertiary prevention. *British Journal of Cancer*, 119(7), 785-792.

Brenner, H., Altenhofen, L., Stock, C., Hoffmeister, M. (2013). Natural history of colorectal adenomas: birth cohort analysis among 3.6 million participants of screening colonoscopy. *Cancer Epidemiology and Prevention Biomarkers*, 22(6), 1043-1051.

Brenner, H., Chen, C. (2018). The colorectal cancer epidemic: challenges and opportunities for primary, secondary and tertiary prevention. *British Journal of Cancer*, 119(7), 785-792.

Brombin, C., Di Serio, C., Rancoita, P. M. (2016). Joint modeling of HIV data in multicenter observational studies: A comparison among different approaches. *Statistical Methods in Medical Research*, 25(6), 2472-2487.

Borges, A., Sousa, I., Castro, L. (2017, June). Joint modelling of longitudinal CEA tumour marker progression and survival data on breast cancer. In *AIP Conference Proceedings* (Vol. 1836, No. 1, p. 020043). AIP Publishing LLC.

Canadian Cancer Society (CCS) (2021). Carcinoembryonic antigen (CEA) https://www.cancer.ca/en/cancer-information/diagnosis-and-treatment/tests-and-procedures/carcinoembryonic-antigen-cea/?region=on

Cekic, S., Aichele, S., Brandmaier, A. M., Köhncke, Y., Ghisletta, P. (2019). A tutorial for joint modeling of longitudinal and time-to-event data in R. *arXiv preprint arXiv*:1909.05661.

Collett, D. (2015). *Modelling survival data in medical research.* CRC press

Colorectal Cancer Alliance (2019). Carcinoembryonic Antigen (CEA) Biomarker. https://www.ccalliance.org/colorectal-cancer-information/biomarkers/biomarkers-cea.

Coppedè, F. (2014). The role of epigenetics in colorectal cancer. *Expert review of gastroenterology hepatology*, 8(8), 935-948.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187-202.

Curran-Everett, D. (2018). Explorations in statistics: the log transformation. *Advances in physiology education*, 42(2), 343-347.

Desmée, S., Mentré, F., Veyrat-Follet, C., Sébastien, B., Guedj, J. (2017). Nonlinear joint models for individual dynamic prediction of risk of death using Hamiltonian Monte Carlo: application to metastatic prostate cancer. *BMC medical research methodology*, 17(1), 1-12

Duffy, M. J., Lamerz, R., Haglund, C., Nicolini, A., Kalousová, M., Holubec, L., Sturgeon, C. (2014). Tumor markers in colorectal cancer, gastric cancer and gastrointestinal stromal cancers: European group on tumor markers 2014 guidelines update. *International journal of cancer*, 134(11), 2513-2522.

Duineveld, L. A. M., Van Asselt, K. M., Bemelman, W. A., Smits, A. B., Tanis, P. J., Van Weert, H. C. P. M., Wind, J.. (2016). Symptomatic and Asymptomatic Colon Cancer Recurrence: A Multicenter Cohort Study. *The Annals of Family Medicine*, 14(3), 215–220. https://doi.org/10.1370/afm.1919

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., ... Bray, F. (2018). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European journal of cancer*, 103, 356-387.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, 136(5), E359-E386

Fletcher, R. H. (1986). Carcinoembryonic antigen. *Annals of internal medicine*, 104(1), 66-73.

Gao, Y., Wang, J., Zhou, Y., Sheng, S., Qian, S. Y., Huo, X. (2018). Evaluation of serum CEA, CA19-9, CA72-4, CA125 and ferritin as diagnostic markers and factors of clinical parameters for colorectal cancer. *Scientific reports*, 8(1), 1-9.

Godhi, S., Godhi, A., Bhat, R., Saluja, S. (2017). Colorectal cancer: postoperative follow-up and surveillance. *Indian Journal of Surgery*, 79(3), 234-237.

Groenwold, R. H. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., Moons, K. G. M.. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, 184(11), 1265–1269. https://doi.org/10.1503/cmaj.110977

Guthrie, J. A. (2002). Colorectal cancer: follow-up and detection of recurrence. *Abdominal Radiology*, 27(5), 570.

Haggar, F. A., Boushey, R. P. (2009). Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4), 191.

Hansen, I. O., Jess, P. (2012). Possible better long-term survival in left versus right-sided colon cancer-a systematic review. *Dan Med J, 59*(6), A4444

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, 247(18), 2543-2546.

Henderson, R., Diggle, P., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4), 465-480.

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

Jeffery, M., Hickey, B. E., Hider, P. N., See, A. M.. (2016). Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database of Systematic Reviews. https://doi.org/10.1002/14651858.cd002200.pub3

Kalbfleisch, J. D., Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley Sons.

Keum, N., & Giovannucci, E. (2019). Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature reviews. Gastroenterology hepatology, 16*(12), 713–732. https://doi.org/10.1038/s41575-019-0189-8

Kuipers, E.J., Grady, W.M., Lieberman, D., Seufferlein, T., Sung, J.J., Boelens, P.G., Van De Velde, C.J.H., Watanabe, T., 2015. Colorectal cancer. *Nature Reviews Disease Primers* 1, 15065.. doi:10.1038/nrdp.2015.65

Kwaan, M. R. (2020). Postoperative CEA and Other Non-traditional Risk Factors for Colon Cancer Recurrence: Findings from Swedish Population-Based Data. *Annals of surgical oncology*, 27(4), 971-972

Lawrence Gould, A., Boye, M. E., Crowther, M. J., Ibrahim, J. G., Quartey, G., Micallef, S., Bois, F. Y. (2015). Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics* in medicine, 34(14), 2181-2195.

Liang, X., Hendryx, M., Qi, L., Lane, D., Luo, J. (2020). Association between prediagnosis depression and mortality among postmenopausal women with colorectal cancer. *Plos one*, 15(12), e0244728.

Macrae, F. A. (2016). Colorectal cancer: Epidemiology, risk factors, and protective factors. *Uptodate com [ažurirano 9. lipnja 2017*

Mant, D., Perera, R., Gray, A., Rose, P., Fuller, A., Corkhill, A., ... Primrose, J. N. (2013). Effect of 3-5 years of scheduled CEA and CT follow-up to detect recurrence of colorectal cancer: *FACS randomized controlled trial.*

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data.* Springer Science & Business Media.

Njagi, E. N., Rizopoulos, D., Molenberghs, G., Dendale, P., Willekens, K.. (2013). A joint survival-longitudinal modelling approach for the dynamic prediction of rehospitalization in telemonitored chronic heart failure patients. *Statistical Modelling, 13*(3), 179–198. https://doi.org/10.1177/1471082x13478880

Park, H.-C., Shin, A., Kim, B.-W., Jung, K.-W., Won, Y.-J., Oh, J. H., Jeong, S.-Y., Yu, C. S., Lee, B. H.. (2013). Data on the Characteristics and the Survival of Korean Patients With Colorectal Cancer From the Korea Central Cancer *Registry. Annals of Coloproctology, 29*(4), 144. https://doi.org/10.3393/ac.2013.29.4.144

Peng, Y., Zhai, Z., Li, Z., Wang, L., Gu, J. (2015). Role of blood tumor markers in predicting metastasis and local recurrence after curative resection of colon cancer. *International journal of clinical and experimental medicine, 8*(1), 982.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., Maintainer, R. (2017). Package 'nlme'. Linear and nonlinear mixed effects models, version, 3(1).

Primeau, A.S.B., (2018, November 30). Cancer Recurrence Statistics. https://www.cancertherapyadvisor.com/home/tools/fact-sheets/cancer-recurrence-statistics/

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics, 67*(3), 819-829.

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

Rizopoulos, D. (2014). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using MCMC. *arXiv preprint arXiv:1404.7625.*

Rizopoulos, D., Rizopoulos, M. D., Imports, M. A. S. S., SystemRequirements, J. A. G. S., Rcpp, L. (2020). Package 'JMbayes'.

Ryuk, J. P., Choi, G. S., Park, J. S., Kim, H. J., Park, S. Y., Yoon, G. S., ... Kwon, Y. C. (2014). Predictive factors and the prognosis of recurrence of colorectal cancer within 2 years after curative resection. *Annals of surgical treatment and research, 86*(3), 143.

Saito, G., Sadahiro, S., Okada, K., Tanaka, A., Suzuki, T., Kamijo, A. (2016). Relation between carcinoembryonic antigen levels in colon cancer tissue and serum carcinoembryonic antigen levels at initial surgery and recurrence. *Oncology, 91*(2), 85-89.

Sargent, D. J., Wieand, H. S., Haller, D. G., Gray, R., Benedetti, J. K., Buyse, M., Labianca, R., Seitz, J. F., O'Callaghan, C. J., Francini, G., Grothey, A., O'Connell, M., Catalano, P. J., Blanke, C. D., Kerr, D., Green, E., Wolmark, N., Andre, T., Goldberg, R. M., De Gramont, A.. (2005). Disease-Free Survival Versus Overall Survival As a Primary End Point for Adjuvant Colon Cancer Studies: Individual Patient Data From 20,898 Patients on 18 Randomized Trials. *Journal of Clinical Oncology, 23*(34), 8664–8670. https://doi.org/10.1200/jco.2005.01.6071

Shinkins, B., Nicholson, B. D., Primrose, J., Perera, R., James, T., Pugh, S., Mant, D. (2017). The diagnostic accuracy of a single CEA blood test in detecting colorectal cancer recurrence: Results from the FACS trial. *PloS one, 12*(3), e0171810.

Simon, K. (2016). Colorectal cancer development and advances in screening. *Clinical interventions in aging, 11*, 967.

Sørensen, C. G., Karlsson, W. K., Pommergaard, H. C., Burcharth, J., Rosenberg, J. (2016). The diagnostic accuracy of carcinoembryonic antigen to detect colorectal cancer recurrence–A systematic review. *International Journal of Surgery, 25*, 134-144.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj, 338*.

Powney, M., Williamson, P., Kirkham, J., Kolamunnage-Dona, R. (2014). A review of the handling of missing longitudinal outcome data in clinical trials. *Trials, 15*(1), 1-11.

Proust-Lima, C., Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biostatistics, 10*(3), 535-549.

Therneau, T. M., Lumley, T. (2015). Package 'survival'. *R Top Doc, 128*(10), 28-33.

Thota, R., Abu Hazeem, M., Kallam, A., Silberstein, P. T., Subbiah, S. (2012). Factors influencing the CEA positivity in patients with colon cancer including race. *Journal of Clinical Oncology*, 30(4_suppl), 409-409.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verslag, New York.

Wang, J. Y., Lu, C. Y., Chu, K. S., Ma, C. J., Wu, D. C., Tsai, H. L., ... Hsieh, J. S. (2007). Prognostic significance of pre-and postoperative serum carcinoembryonic antigen levels in patients with colorectal cancer. *European Surgical Research, 39*(4), 245-250.

Wieldraaijer, T., Bruin, P., Duineveld, L. A., Tanis, P. J., Smits, A. B., van Weert, H. C., Wind, J. (2018). Clinical pattern of recurrent disease during the follow-up of rectal carcinoma. *Digestive surgery, 35*(1), 35-41.

Wu, T., Mo, Y., Wu, C. (2020). Prognostic values of CEA, CA19-9, and CA72-4 in patients with stages I-III colorectal cancer. *International journal of clinical and experimental pathology, 13*(7), 1608–1614.

Wulfsohn, M. S., Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, 330-339.

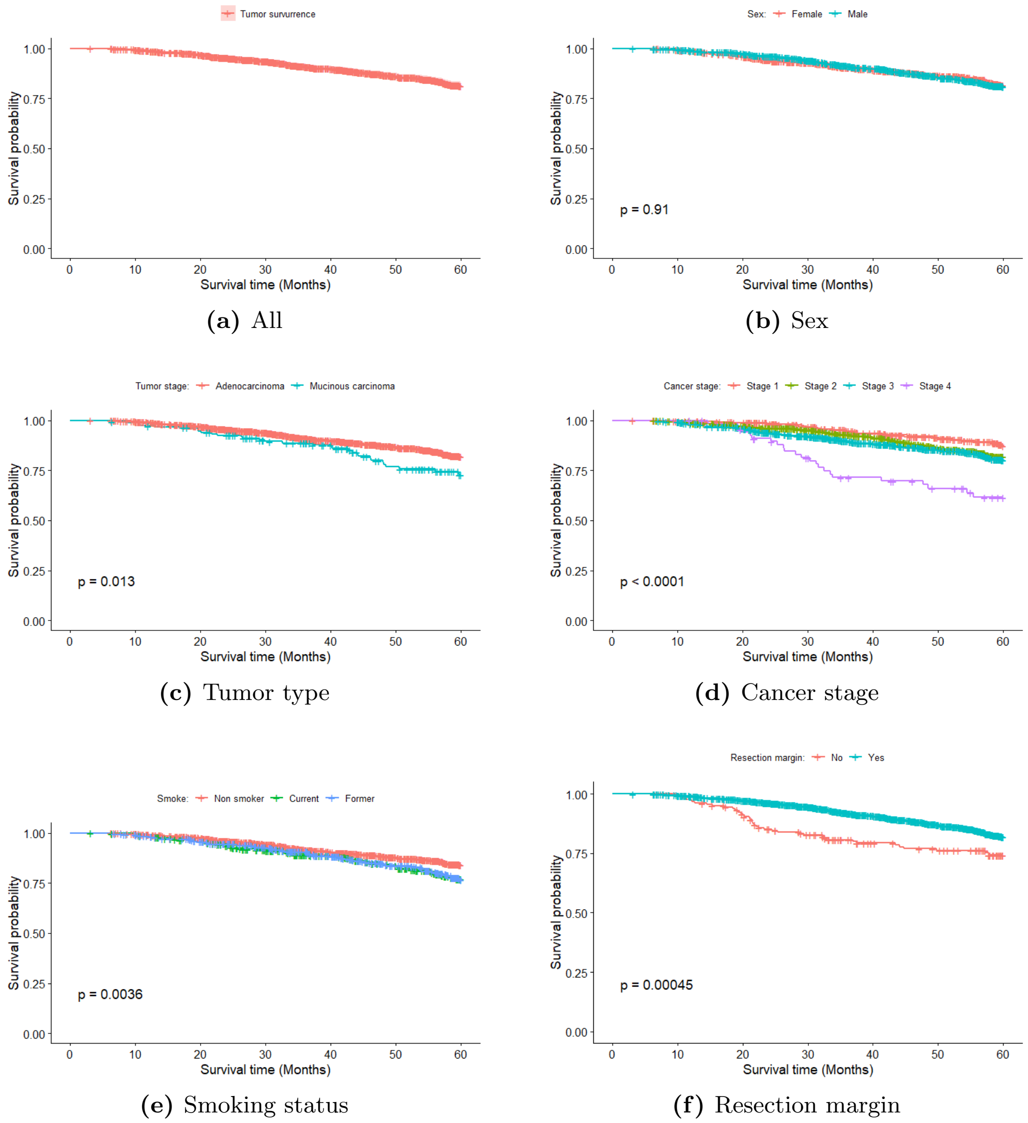
Ye, W., Lin, X., Taylor, J. M. (2008). Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach. *Biometrics, 64*(4), 1238-1246.

Yun, H. R., Lee, L. J., Park, J. H., Cho, Y. K., Cho, Y. B., Lee, W. Y., ... Yun, S. H. (2008). Local recurrence after curative resection in patients with colon and rectal cancers. *International journal of colorectal disease, 23*(11), 1081-1087.

Zuyderland (2021). About Zuyderland. https://www.zuyderland.nl/english/

A Appendix

A.1 Kaplan-Meier curves



(a) All

(b) Sex

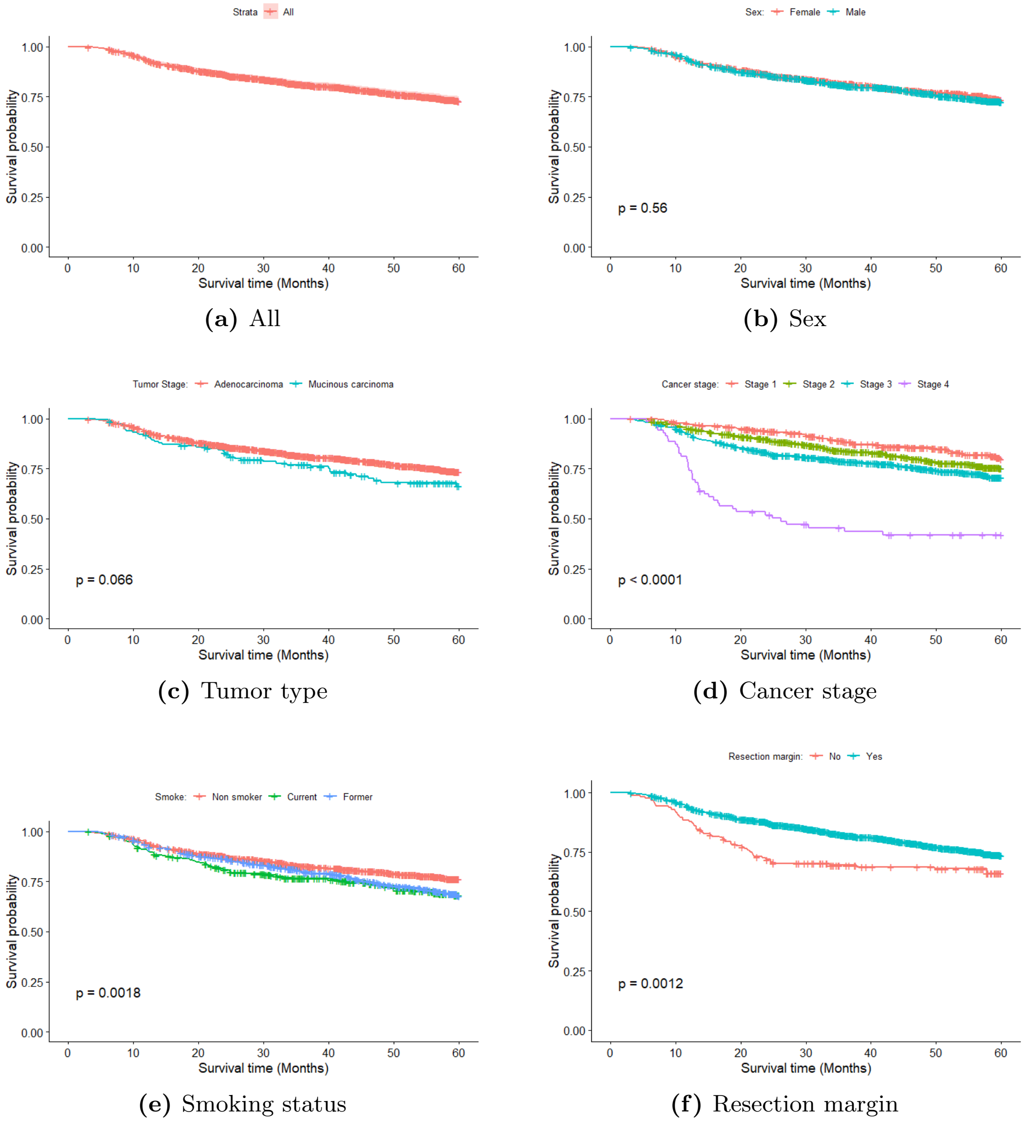
(c) Tumor type

(d) Cancer stage

(e) Smoking status

(f) Resection margin

Figure A.1: Overall and variable specific survival curves (*Death*). For sex (b), Tumor type (c), Cancer stage (d), Smoking status (e), and resection margin (f). The *p*-value was obtained using the log-rank test



(a) All

(b) Sex

(c) Tumor type

(d) Cancer stage

(e) Smoking status

(f) Resection margin

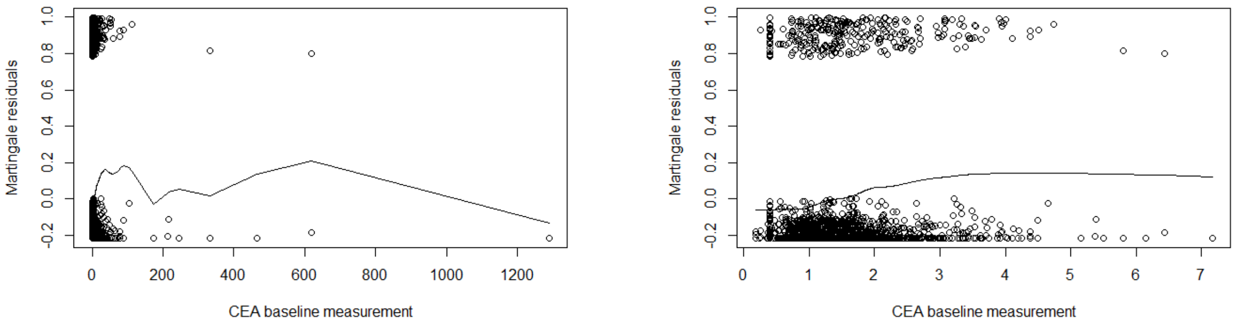
Figure A.2: Overall and variable specific survival curves (*Tumor recurrence or Death*). For sex (b), Tumor type (c), Cancer stage (d), Smoke (e), and resection margin (f). The p-value was obtained using the log-rank test

A.2 Linear mixed model type three F test

Table A.1: Type 3 Tests of Fixed Effects

Variables	df	F-value	p-value
(Intercept)	1	846.7362	<.0001
Age	1	6.762	0.0094
Smoking	2	46.7452	<.0001
Cancer stage	3	4.4428	0.0041
Time	1	187.2354	<.0001
Age:Time	1	6.4512	0.0111

A.3 Cox proportional hazard model



(a)

(b)

Figure A.3: Functional form of pre-surgery CEA measurements based on Martingale Residual Plots before log transforming (Figure A.3a) and after log transforming (Figure A.3b).

Table A.2: Schoenfeld's test for the proportional hazard assumption in the Cox Model

	chi square	Df	P value
Age	3.4100	1	0.065
Sex	0.6870	1	0.407
Smoking	0.7280	2	0.695
Tumor type	0.0004	1	0.985
Tumor stage	4.1700	3	0.244
Resection Margin	5.1100	1	0.024
Log CEA baseline	1.4800	1	0.224
GLOBAL	13.0000	10	0.225

Table A.3: Schoenfeld's test for the proportional hazard assumption in the Cox Model stratified by resection margin

	chi square	Df	P value
Age	3.553	1	0.059
Sex	0.959	1	0.327
Smoking	0.850	2	0.654
Tumor type	0.066	1	0.797
Tumor stage	2.985	3	0.394
Log CEA baseline	0.628	1	0.428
GLOBAL	7.687	9	0.566

Table A.4: The overall effect of each covariate in the Cox PH model fitted to the training dataset

Variables	**Likelihood Ratio Test**	**Df**	**p value**
Age	5.1886	1	0.0227
Sex	0.1441	1	0.7042
Smoking	2.9359	2	0.2304
type	0.1325	1	0.7159
tumor stage	71.1416	3	<0.0001
Log CEA baseline	15.549	1	<0.0001

A.4 Joint models

Table A.5: AIC, BIC, Log-likelihood values for the four joint models fitted to the training dataset. The last three columns indicates the likelihood ratio test statistics, degrees of freedom and *p*-values for testing the null model (I) against model (II), (III), and (IV), respectively

Joint model parameterization	**AIC**	**BIC**	**loglikelihood**	**LRT**	**df**	**P value**
I: Current value (No baseline covariates)	8043.46	8092.29	-4012.73			
II: Current value	7868.92	8020.86	-3906.46	212.54	19	<0.0001
III: Slope value	7890.68	8042.63	-3917.34	190.77	19	<0.0001
IV: Current value plus slope	7849.13	8006.50	-3895.56	234.33	20	<0.0001

Model I: no baseline covariates in both survival and longitudinal submodels

Model II, III, IV includes the baseline covariates in both survival and longitudinal submodels

Table A.6: Parameter estimates and 95% confidence intervals for the "current value" parameterization results fitted to the training dataset

Variable	**Estimate**	95% CI	**SE**	**p-value**	RE
Longitudinal Sub-model					
Intercept	0.9249	(0.8625; 0.9873)	0.0318	<0.0001	2.5217
Age (above 75 years)	0.0788	(0.0185; 0.1391)	0.0308	0.0104	1.082
Smoking					
Current smokers	0.3305	(0.2626; 0.3985)	0.0347	<0.0001	1.3917
Former smokers	0.1231	(0.0626; 0.1837)	0.0309	0.0001	1.1311
Tumor stage					
Two	0.0388	(-0.0323; 0.1100)	0.0363	0.2845	1.0396
Three	0.0661	(-0.0025; 0.1346)	0.035	0.0589	1.0683
Four	0.2665	(0.1157; 0.4173)	0.0769	0.0005	1.3053
Time	0.0061	(0.0052; 0.0069)	0.0004	<0.0001	1.0061
Age:Time	0.0026	(0.0007; 0.0045)	0.001	0.0081	1.0026
Variable	**Estimate**	95% CI	**SE**	**p-value**	**HR 95% CI**
Survival Sub-model					
Intercept	-7.3383	(-7.9490; -6.7276)	0.3116	<0.0001	-
Age (above 75 years)	0.3520	(0.0736; 0.6304)	0.1421	0.0132	1.4219 (1.0763; 1.8784)
Sex (male)	-0.2175	(-0.4665; 0.0315)	0.127	0.0869	0.8045 (0.6272; 1.0320)
Smoking					
Current smokers	0.1255	(-0.1860; 0.4370)	0.1589	0.4298	1.1337 (0.8303; 1.5480)
Former smokers	0.1536	(-0.1398; 0.4471)	0.1497	0.3048	1.1661 (0.8695; 1.5638)
Tumor type (mucinous carcinoma)	-0.0539	(-0.5229; 0.4151)	0.2393	0.8217	0.9475 (0.5928; 1.5146)
Tumor stage					
Two	0.2376	(-0.2139; 0.6892)	0.2304	0.3022	1.2683 (0.8075; 1.9921)
Three	0.7291	(0.3072; 1.1510)	0.2153	0.0007	2.0732 (1.3596; 3.1614)
Four	2.0983	(1.5661; 2.6305)	0.2715	<0.0001	8.1526 (4.7881; 13.8811)
Log CEA baseline	0.1314	(0.0066; 0.2562)	0.0637	0.0391	1.1404 (1.0066; 1.2920)
Current value association	0.9492	(0.8444; 1.0540)	0.0535	<0.0001	2.5836 (2.3266; 2.8691)

Table A.7: Parameter estimates and 95% confidence intervals for the "current slope" joint model parameterization results fitted to the training dataset

Variable	**Estimate**	**lower**	**SE**	**p-value**	RE
Longitudinal Sub-model					
Intercept	0.9155	(0.8540; 0.9769)	0.0313	<0.0001	2.4979
Age (above 75 years)	0.0790	(0.0214; 0.1367)	0.0294	0.0072	1.0822
Smoking					
Current smokers	0.3287	(0.2621; 0.3954)	0.0340	<0.0001	1.3892
Former smokers	0.1273	(0.0702; 0.1843)	0.0291	<0.0001	1.1357
Tumor stage					
Two	0.0450	(-0.0239; 0.1139)	0.0352	0.2005	1.0460
Three	0.0666	(-0.0010; 0.1342)	0.0345	0.0536	1.0688
Four	0.2434	(0.0940; 0.3928)	0.0762	0.0014	1.2756
Time	0.0068	(0.0059; 0.0077)	0.0005	<0.0001	1.0069
Age (above 75 years):months	0.0026	(0.0007; 0.0046)	0.0010	0.0082	1.0026
Variable	**Estimate**	**95% CI**	**SE**	**p-value**	**HR (95% CI)**
Survival Sub-model					
Intercept	-7.0007	(-7.6252; -6.3762)	0.3186	<0.0001	-
Age (above 75 years)	0.2538	(-0.0294; 0.5370)	0.1445	0.0790	1.2889 (0.9710; 1.7109)
Sex (male)	-0.2507	(-0.5072; 0.0058)	0.1309	0.0554	0.7783 (0.6022; 1.0058)
Smoking					
Current smokers	0.4087	(0.0921; 0.7254)	0.1616	0.0114	1.5049 (1.0964; 2.0656)
Former smokers	0.3062	(0.0081; 0.6043)	0.1521	0.0441	1.3583 (1.0081; 1.8300)
Tumor type (Mucinous carcinoma)	0.0644	(-0.4082; 0.5371)	0.2411	0.7894	1.0665 (0.6648; 1.7110)
Tumor stage					
Two	0.2557	(-0.1956; 0.7070)	0.2303	0.2668	1.2914 (0.8223; 2.0279)
Three	0.6818	(0.2583; 1.1054)	0.2161	0.0016	1.9775 (1.2947; 3.0204)
Four	1.9908	(1.4466; 2.5350)	0.2777	<0.0001	7.3213 (4.2486; 12.6165)
Log CEA baseline	0.2049	(0.0864; 0.3234)	0.0604	0.0007	1.2274 (1.0903; 1.3818)
Slope association	30.8016	(26.6870; 34.9162)	2.0993	<0.0001	

Table A.8: The overall effect of each covariate in the "current value plus slope" joint model parameterization fitted to the training dataset

	Likelihood Ratio Test	**Df**	**p value**
Age	4.9000	1	0.0237
Sex	3.4900	1	0.1749
Smoking	4.7100	2	0.0300
Tumor type	0.1100	1	0.9984
tumor stage	62.3100	3	<0.0001
Log CEA baseline	5.5100	1	0.0189
Current value association	16.3200	1	<0.0001
Slope association	46.8900	1	<0.0001

Table A.9: Internal validity results showing the area under the curves (AUCs) computed from test data, under various joint model formulations

Joint model parameterization	Time window Δt	Information used up time t	time point	AUC (t)
I: Current value **No baseline covariates**	3	7	10	0.6324
		17	20	0.5446
		27	30	0.7042
		37	40	0.7310
		47	50	0.7147
		57	60	0.5525
	6	4	10	0.6452
		14	20	0.6009
		24	30	0.5104
		34	40	0.7549
		44	50	0.5508
		54	60	0.4676
II: Current value	3	7	10	0.7001
		17	20	0.6927
		27	30	0.7748
		37	40	0.6825
		47	50	0.7136
		57	60	0.5938
	6	4	10	0.6266
		14	20	0.5999
		24	30	0.6500
		34	40	0.4605
		44	50	0.6221
		54	60	0.4909
III: Current slope	3	7	10	0.7280
		17	20	0.7163
		27	30	0.6635
		37	40	0.7035
		47	50	0.7956
		57	60	0.6381
	6	4	10	0.5889
		14	20	0.6185
		24	30	0.7364
		34	40	0.3044
		44	50	0.7032
		54	60	0.5053
IV: Current value plus slope	3	7	10	0.7062
		17	20	0.7407
		27	30	0.7072
		37	40	0.6982
		47	50	0.7723
		57	60	0.5920
	6	4	10	0.6111
		14	20	0.6316
		24	30	0.6915
		34	40	0.3933
		44	50	0.6554
		54	60	0.4839

Model I: no baseline covariates in both survival and longitudinal submodels

Model II, III, IV includes the baseline covariates in both survival and longitudinal submodels

A.5 More equations

The log baseline hazard risk function log $h_0(t)$ is expressed into B-splines as follows.

$$log\ h_0(t) = k_0 + \sum_{q=1}^{Q} k_q B_q(t, v) \tag{A.18}$$

Where k is a vector of the spline coefficient and $B_q(t, v)$ represent the q^{th} basis function of a B-spline with knots $v_1, \cdots, v_q$.

$$\begin{aligned} h_i(t) = h_0(t)exp[\beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \sum_{s=1}^{2} \beta_{3s} Smoke_s + \beta_4 Tumor\_Type_i \\ + \sum_{s=2}^{4} \beta_{5s} Tumor\_Stage_s + \beta_6 Pre\_Surgery\_log\_CEA + \alpha m_i(t)\{max(t-1, 0)\}] \end{aligned} \tag{A.19}$$

Equation A.19 is the lagged effect joint model formulation, which assumes the the risk of tumor recurrence depends on the previous log CEA value earlier the current log CEA value.

A.6 R Code

```
#clear the space
rm(list=ls())

#Load the required packages
library(rjags)
require(rstan)
library(pacman)
p_load(readxl, ggplot2, dplyr,survival,JM,survival,survminer,janitor,ggplot2,epicalc, reshape2,lattice)


# Load the data
df <- read_excel("CEA_masters_20210324.xlsx")
dim(df);str(df)

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
#                       Functions for extracting the parameters
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
# LME

print.lmm.wald<-function(mod.lmm)
{
  c.tab<-coef(summary(mod.lmm))
  new.tab<-data.frame(round(c.tab[,1],4),round(c.tab[,2], 4),round(c.tab[,1]-1.96*c.tab[,2], 4),round(c.tab[,1]+1.96*c.tab[,2],4),
  round(c.tab[,5],4),round(exp(c.tab[,1]),4))
  names(new.tab)<-c("Estimates","SE","L95", "U95","p-value","RE %")
  return(new.tab)
}

# Cox model

print.HRCIs<-function(mod.coxph)
{
  log_hazard <-coef(summary(mod.coxph))[,1]
  L95 <- coef(summary(mod.coxph))[,1]-1.96*coef(summary(mod.coxph))[,3]
  U95 <- coef(summary(mod.coxph))[,1]+1.96*coef(summary(mod.coxph))[,3]
  SE <- coef(summary(mod.coxph))[,3]
  HR<-coef(summary(mod.coxph))[,2]
  H_L95<-summary(mod.coxph)$conf.int[,3]
  H_U95<-summary(mod.coxph)$conf.int[,4]
  p.val<-coef(summary(mod.coxph))[,5]
  new.tab<-data.frame(round(log_hazard, 4), round(L95, 4), round(U95, 4),round(SE, 4),round(p.val,4),round(HR, 4), round(H_L95, 4), round(H_U95, 4))
  names(new.tab)<-c("log_hazard","L95","U95","SE","P.val","HR", "H_L95", "H_U95")
  row.names(new.tab)<-t(t(row.names(coef(summary(mod.coxph)))))
  return(new.tab)
}

# Joint Model
print.joint.lda<-function(mod.joint.lda)
{
  Estimates <- coef(summary(mod.joint.lda))$Longitudinal[,1]
  L95 <- coef(summary(mod.joint.lda))$Longitudinal[,1] - 1.96*coef(summary(mod.joint.lda))$Longitudinal[,2]
  U95 <- coef(summary(mod.joint.lda))$Longitudinal[,1] + 1.96*coef(summary(mod.joint.lda))$Longitudinal[,2]
  SE <- coef(summary(mod.joint.lda))$Longitudinal[,2]
  p.val <- coef(summary(mod.joint.lda))$Longitudinal[,4]
  RE <- exp(coef(summary(mod.joint.lda))$Longitudinal[,1])
  new.tab<-data.frame(round(Estimates, 4), round(L95, 4), round(U95, 4),round(SE, 4),round(p.val,4),round(RE, 4))
  names(new.tab)<-c("Estimates","L95","U95","SE","P.val","RE")
  return(new.tab)
}
```

```
print.joint.event<-function(mod.joint.event)
{
  Estimates <- coef(summary(mod.joint.event))$Event[,1]
  L95 <- coef(summary(mod.joint.event))$Event[,1] - 1.96*coef(summary(mod.joint.event))$Event[,2]
  U95 <- coef(summary(mod.joint.event))$Event[,1] + 1.96*coef(summary(mod.joint.event))$Event[,2]
  SE <- coef(summary(mod.joint.event))$Event[,2]
  p.val <- coef(summary(mod.joint.event))$Event[,4]
  HR <- exp(coef(summary(mod.joint.event))$Event[,1])
  LHR_95 <- exp(coef(summary(mod.joint.event))$Event[,1] - 1.96*coef(summary(mod.joint.event))$Event[,2])
  UHR_95 <- exp(coef(summary(mod.joint.event))$Event[,1] + 1.96*coef(summary(mod.joint.event))$Event[,2])
  new.tab<-data.frame(round(Estimates, 4), round(L95, 4), round(U95, 4),round(SE, 4),round(p.val,4),round(HR, 4),round(LHR_95, 4), round(UHR_95, 4))
  names(new.tab)<-c("Estimates","L95","U95","SE","P.val","HR","LHR_95","UHR_95")
  return(new.tab)
}

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
#                                 Data management                                                                        #
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

## ------- Recoding varibales ------
#------------------------------------------------------------------------------------------------------------------------
## Tumor type
df$Tumor_cat <- as.factor(ifelse(df$Tumor == 0, "Adenocarcinoma",ifelse(df$Tumor == 1,"Mucinous carcinoma", "Other")))
tab1(df$Tumor_cat)

## TResection merging
df$ResectionMargeFree_cat <- as.factor(ifelse(df$ResectionMargeFree == 1, "Yes","No"))
tab1(df$ResectionMargeFree_cat)

## Tumor stage
df$Stadium_cat1 <- as.factor(ifelse(df$Stadium_cat == "T1/T2 N0 M0","stage 1",ifelse(df$Stadium_cat == "T3/T4 N0 M0","stage 2",
                                                                                     ifelse(df$Stadium_cat == "T1-4 N1-3 M0","stage 3","stage 4"))))

table(df$Stadium_cat)
table(df$Stadium_cat1)

df$Stadium_cat2 <- as.factor(ifelse((df$Stadium_cat == "T1/T2 N0 M0" | df$Stadium_cat == "T3/T4 N0 M0"),"stage 1 and 2","stage 3 and 4"))
table(df$Stadium_cat)
table(df$Stadium_cat2)

# Age category
#https://www.cancer.net/cancer-types/colorectal-cancer/risk-factors-and-prevention
df$Age_cat <- as.factor(ifelse(df$AgeOK <= 75, "0", "1"))
tab1(df$Age_cat)


#------------------------------------------------------------------------------------------------------------------------

#Basic statistics of 'CEA Outcome 1 and CEA Outcome 1'
summary(df$CeaOutcome1)
summary(df$CeaOutcome2)

CeaOutcome1 <- df$CeaOutcome1
tab1(df$CeaOutcome1, graph = T)

CeaOutcome1.1 <- ifelse((df$CeaOutcome1>0 & df$CeaOutcome1 <= 20),df$CeaOutcome1,NA)

hist(CeaOutcome1.1, xlim = c(0,20), nclass = 20)

summary(CeaOutcome1.1)
```

```
hist(df$CeaBaseline, xlim = c(0,20), nclass = 20)

#**********************************************************************************************************
# Time difference (in days study follow period 5yrs (we drop patients with more than 5 years followup))
#**********************************************************************************************************

head(df[,c("Date_FU_Recurrence","Date_FU_Surv_Rec","Date_FU_Surv")],20)

df$time_rec  =  as.numeric(as.Date(df$Date_FU_Recurrence)- as.Date(df$DateFinalOperation)) # Tumor reccurrence
df$time_DFS  = as.numeric(as.Date(df$Date_FU_Surv_Rec) - as.Date(df$DateFinalOperation)) # Tumor reccurence or Death
df$time_surv = as.numeric(as.Date(df$Date_FU_Surv) - as.Date(df$DateFinalOperation)) # Death

df[,c("Id","DateRecurrence","Laatste_FU_datum","LastSeenHospital",
      "Date_FU_Recurrence","DateMortality","Date_FU_Surv","Date_FU_Surv_Rec")]

#*******************************************************************************************
#Number of patients with DFS ( Disease Free Survival (event is death or reccurence)) longer than 5 year:
#**********************************************************************************************************

length(which(df$time_DFS>1825))
length(which(df$time_DFS<=1825))

#Replace FU times longer than 5 year (1826 days) with 1826 and censor the event variable
df$day_DFS_max1826 = df$time_DFS
df$cen_dfs_1826 = df$cen_dfs

for(i in 1:dim(df)[1]){
  if(df$time_DFS[i] > 1825){df$day_DFS_max1826[i] <- 1826}
  if(df$time_DFS[i] > 1825){df$cen_dfs_1826[i] <- 0}
}

table(df$cen_dfs[df$time_DFS <= 1825])
table(df$cen_dfs_1826)

#*******************************************************************************************
#Number of patients with reccurence longer than 5 year:
#*******************************************************************************************
length(which(df$time_rec>1825))
length(which(df$time_rec<=1825))

#Replace Recurrence times longer than 5 year (1826 days) with 1826 and censor the event variable
df$day_rec_max1826 = df$time_rec
df$cen_rec_1826 = df$cen_rec

for(i in 1:dim(df)[1]){
  if(df$time_rec[i] > 1825){df$day_rec_max1826[i] <- 1826}
  if(df$time_rec[i] > 1825){df$cen_rec_1826[i] <- 0}
}

table(df$cen_rec)
table(df$cen_rec[df$time_rec <= 1825])
table(df$cen_rec_1826)



#*******************************************************************************************
#Number of patients with event is death longer than 5 year:
#**********************************************************************************************************
length(which(df$time_surv>1825))
length(which(df$time_surv<=1825))
```

```
#Replace FU times longer than 5 year (1826 days) with 1826 and censor the death event variable
df$day_surv_max1826 = df$time_surv
df$cen_surv_1826 = df$cen_surv

for(i in 1:dim(df)[1]){
  if(df$time_surv[i] > 1825){df$day_surv_max1826[i] <- 1826}
  if(df$time_surv[i] > 1825){df$cen_surv_1826[i] <- 0}
}

table(df$cen_surv[df$time_surv <= 1825])
table(df$cen_surv_1826)



#------------------------------------------
# Change date difference from days to months
#------------------------------------------
df$month_rec_max60 = 12*(df$day_rec_max1826/365.25)
df$month_DFS_max60  = 12*(df$day_DFS_max1826/365.25)
df$month_surv_max60  = 12*(df$day_surv_max1826/365.25)

summary(df[,c("month_rec_max60","month_DFS_max60","month_surv_max60")])

#----------------------------------------------------------------------------------------
# CEA measurement taken => six months before surgery (code cea at baseline to be missing)
#----------------------------------------------------------------------------------------

head(df[,c("DateFinalOperation","DateCeaBaseline","CeaBaseline")],20)

df$time_difference <- as.numeric(as.Date(df$DateFinalOperation) - as.Date(df$DateCeaBaseline))
summary(df$time_difference)

df$time_diff_months <- 12*(df$time_difference/365.25)
summary(df$time_diff_months)

df$CeaBaseline1 <- ifelse(df$time_diff_months < 6,df$CeaBaseline,NA)
summary(df$CeaBaseline1)

hist(df$CeaBaseline1, xlim = c(0,20), breaks = 10000)

length(which(is.na(df$CeaBaseline1))) # 282
length(which(!is.na(df$CeaBaseline1))) # 1835

#------------------------------------------------------------------------
# Imputing the missing covariate (Smoking, Tumor type, Resection merging)
#-------------------------------------------------------------------------
library(pacman)
p_load(tidyverse,haven,mice,sjPlot,sjmisc,VIM)

#coding other to missing (Tumor type)
table(df$Tumor)
df$Tumor_1 <- ifelse(df$Tumor == 3,NA,df$Tumor)
table(df$Tumor_1)

## plot of missingness (baseline characteristics)
VIM::aggr(df[,c("CeaBaseline1","Sex_cat","Smoke_cat","Tumor_1","Stadium_cat","ResectionMargeFree_cat")])


dat_imp <- df[,c("Id","CeaBaseline1","Smoke","Tumor_1","ResectionMargeFree")]

dat_imp <- dat_imp %>%
  mutate_at(c("Smoke","Tumor_1","ResectionMargeFree"), as.factor)

set.seed(12345)
dat_imp_m <- mice(dat_imp, m = 5, method = c("","pmm","polyreg","logreg","logreg"), maxit = 0)
summary(dat_imp_m)

dat_imp_final <- complete(dat_imp_m, 1)
summary(dat_imp_final)

summary(df$CeaBaseline1)
summary(dat_imp_final$CeaBaseline1)

table(dat_imp_final$Smoke)
table(dat_imp_final$Tumor_1)
table(dat_imp_final$ResectionMargeFree)

head(dat_imp_final,10)

colnames(dat_imp_final) <-  c("Id","CeaBaseline_imp","Smoke_imp","Tumor_imp","ResectionMargeFree_imp")
head(dat_imp_final,10)

# merge original data with imputed data
df_imp  <- merge(df,dat_imp_final,by = "Id", all =T)
dim(df_imp)

df <- df_imp

# ---------------- log transform the baseline CEA value
df$CeaBaseline_imp_log <- log(df$CeaBaseline_imp+1)


df$CeaBaseline_cat <- as.factor(ifelse(df$CeaBaseline_imp >=  2.660, "Yes","No"))
tab1(df$CeaBaseline_cat)

#------------------------------
df$Stadium_two <- as.numeric(df$Stadium_cat1 == "stage 2")
df$Stadium_three<- as.numeric(df$Stadium_cat1 == "stage 3")
df$Stadium_four <- as.numeric(df$Stadium_cat1 == "stage 4")

df$Smoke_1 <- as.numeric(df$Smoke_imp == 1)
df$Smoke_2 <- as.numeric(df$Smoke_imp == 2)

table(df$Stadium_cat1)
table(df$Smoke_imp)

table(df$Stadium_two);table(df$Stadium_three);table(df$Stadium_four)
table(df$Smoke_1);table(df$Smoke_2)

df$Smoke_imp_cat <- ifelse(df$Smoke_imp == 1 | df$Smoke_imp == 2, "smoking","non_smokers")
table(df$Smoke_imp_cat)


#--------------------------------------------------------------------------------------------------------------------------
########################################################
# preparing the data from wide formart to long format
#######################################################
#--------------------------------------------------------------------------------------------------------------------------

library(reshape2)

# Make list of clinical variables that should be included in wide dataframe
```

```
df_wide <- df[,c("Id","CeaBaseline","DateCeaBaseline","Sex","Sex_cat","AgeOK","Age_cat","Smoke","Smoke_cat","Smoke_imp","Smoke_imp_cat",
                 "DateOperation","StadiumCancer","Stadium_cat1","Stadium_cat2","AdjuvChemo","Neoadj_group",
                 "Tumor","Tumor_cat","Tumor_imp","CeaBaseline1","CeaBaseline_imp","CeaBaseline_imp_log","CeaBaseline_cat",
                 "ResectionMargeFree","ResectionMargeFree_cat","ResectionMargeFree_imp","DateFinalOperation",
                 "day_DFS_max1826","month_DFS_max60","cen_dfs_1826",
                 "time_rec","month_rec_max60","cen_rec_1826","cen_surv_1826","month_surv_max60",
                 "Stadium_two","Stadium_three","Stadium_four","Smoke_1","Smoke_2",
                 "CeaDate1","CeaDate2","CeaDate3","CeaDate4","CeaDate5","CeaDate6","CeaDate7","CeaDate8","CeaDate9","CeaDate10",
                 "CeaDate11","CeaDate12","CeaDate13","CeaDate14","CeaDate15","CeaDate16","CeaDate17","CeaDate18","CeaDate19","CeaDate20",
                 "CeaDate21","CeaDate22","CeaDate23","CeaDate24","CeaDate25","CeaDate26","CeaDate27","CeaDate28","CeaDate29","CeaDate30",
                 "CeaDate31","CeaDate32","CeaDate33","CeaDate34","CeaDate35","CeaDate36","CeaDate37","CeaDate38","CeaDate39","CeaDate40",
                 "CeaOutcome1","CeaOutcome2","CeaOutcome3","CeaOutcome4","CeaOutcome5","CeaOutcome6","CeaOutcome7","CeaOutcome8",
                 "CeaOutcome9","CeaOutcome10","CeaOutcome11","CeaOutcome12","CeaOutcome13","CeaOutcome14","CeaOutcome15","CeaOutcome16",
                 "CeaOutcome17","CeaOutcome18","CeaOutcome19","CeaOutcome20","CeaOutcome21","CeaOutcome22","CeaOutcome23","CeaOutcome24",
                 "CeaOutcome25","CeaOutcome26","CeaOutcome27","CeaOutcome28","CeaOutcome29","CeaOutcome30","CeaOutcome31","CeaOutcome32",
                 "CeaOutcome33","CeaOutcome34","CeaOutcome35","CeaOutcome36","CeaOutcome37","CeaOutcome38","CeaOutcome39","CeaOutcome40"
)]
str(df_wide)
df_wide<-as.data.frame(df_wide)

# Make list of clinical variables that should be included in long dataframe
df_long <- reshape(df_wide, idvar = "Id",
                   varying = list(c("CeaDate1","CeaDate2","CeaDate3","CeaDate4","CeaDate5","CeaDate6","CeaDate7","CeaDate8","CeaDate9","CeaDate10",
                                    "CeaDate11","CeaDate12","CeaDate13","CeaDate14","CeaDate15","CeaDate16","CeaDate17","CeaDate18","CeaDate19",
                                    "CeaDate20","CeaDate21","CeaDate22","CeaDate23","CeaDate24","CeaDate25","CeaDate26","CeaDate27","CeaDate28",
                                    "CeaDate29","CeaDate30","CeaDate31","CeaDate32","CeaDate33","CeaDate34","CeaDate35","CeaDate36","CeaDate37",
                                    "CeaDate38","CeaDate39","CeaDate40"),
                                  c("CeaOutcome1","CeaOutcome2","CeaOutcome3","CeaOutcome4","CeaOutcome5","CeaOutcome6","CeaOutcome7",
                                  "CeaOutcome8","CeaOutcome9","CeaOutcome10","CeaOutcome11","CeaOutcome12","CeaOutcome13","CeaOutcome14",
                                  "CeaOutcome15","CeaOutcome16","CeaOutcome17","CeaOutcome18","CeaOutcome19","CeaOutcome20","CeaOutcome21",
                                  "CeaOutcome22","CeaOutcome23","CeaOutcome24","CeaOutcome25","CeaOutcome26","CeaOutcome27","CeaOutcome28",
                                  "CeaOutcome29","CeaOutcome30","CeaOutcome31","CeaOutcome32","CeaOutcome33","CeaOutcome34","CeaOutcome35",
                                  "CeaOutcome36","CeaOutcome37","CeaOutcome38","CeaOutcome39","CeaOutcome40")),
                   v.names = c("CEADATE", "Cea_measure"), direction = "long")



#Check for the duplicate
#which(duplicated(df_long[,c("Id")]))
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
# sorting the data by ID

df_long1 <-  df_long %>%
  arrange(Id, CEADATE)

# Remove missing CEA measurement dates
df_long2 <- df_long1[complete.cases(df_long1[ ,"CEADATE"]),]

# Remove CEA measurement before date of the finaloperation

df_long2$cea_datebefore_operation <- ifelse(df_long2$CEADATE < df_long2$DateFinalOperation, 1,0)

tab1(df_long2$cea_datebefore_operation)

df_long2 <- df_long2[which(df_long2$cea_datebefore_operation == 0),]
dim(df_long2)


# Compute duration from DateOperation till date CEA measurement

df_long2$time_cea = as.Date(df_long2$CEADATE) - as.Date(df_long2$DateFinalOperation)
```

```
head(df_long2[, c("DateFinalOperation", "CEADATE", "time_cea")], 20)


#Change date difference in time_cea to numeric then to months

df_long2$time_cea_num = as.numeric(df_long2$time_cea)
df_long2$time_cea_month =12*(df_long2$time_cea_num/365.25)

summary(df_long2$time_cea_month)


head(df_long2[, c("DateFinalOperation", "CEADATE", "time_cea","time_cea_month")], 20)

# Remove CEA measurements taken after a Tumor recurrence or after maximum FU of 60 months (after operation)
df_long2$cea_dateafter_rec_60 <- ifelse(df_long2$time_cea_month <  df_long2$month_DFS_max60, 1,0)
tab1(df_long2$cea_dateafter_rec_60)

head(df_long2[, c("Id","DateFinalOperation", "CEADATE", "time_cea","time_cea_month","month_DFS_max60","cea_dateafter_rec_60")], 20)

df_long2.2 <- df_long2[which(df_long2$cea_dateafter_rec_60 == 1),]
dim(df_long2.2)

head(df_long2.2[, c("DateFinalOperation", "CEADATE", "time_cea","time_cea_month","month_DFS_max60")], 20)
summary(df_long2.2$time_cea_month)

df_long2.2 <- df_long2.2[which(df_long2.2$time_cea_month > 0),]
dim(df_long2.2)

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

#######################################################################
# Log transforming (the CEA outcome to reduce the skewness of the data)
#######################################################################
df_long2.2$Cea_measure1 <- ifelse((df_long2.2$Cea_measure == -0.5 & df_long2.2$Id == 1760),0,df_long2.2$Cea_measure)
#--------------------------------------------------------
df_long2.2$Cea_measure_ln <- log(df_long2.2$Cea_measure1+1)
#--------------------------------------------------------
#head(df_long2.2[,c("Cea_measure_ln","Cea_measure","Cea_measure1")],20)
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

#######################################################################################################
#          No of individuals with a follow up data.
#          Make sure the number of patients in time to event data (wide format) is the same with the
#          longitindal data (long format)
#######################################################################################################

df_long2.3 <- df_long2.2[,c("Id","Cea_measure","Cea_measure_ln","time_cea_month","AgeOK","Age_cat","Sex_cat","Smoke_imp","Tumor_imp",
                            "Stadium_cat2","Stadium_cat1","ResectionMargeFree_imp","CeaBaseline_imp_log","month_rec_max60",
                            "cen_rec_1826","Stadium_two","Stadium_three","Stadium_four","Smoke_1","Smoke_2","Smoke_imp_cat")]

df_long2.3 <- df_long2.3 %>%
  mutate_at(c("Smoke_imp","Tumor_imp","ResectionMargeFree_imp","Stadium_cat1","Stadium_cat2","Smoke_imp_cat","Sex_cat",
  "Stadium_two","Stadium_three", "Stadium_four","Smoke_1","Smoke_2"), as.factor)

dim(df_long2.3)

df_long2.4 <- df_long2.3[complete.cases(df_long2.3), ]
dim(df_long2.4)


long <- df_long2.4[!duplicated(df_long2.4$Id), ]
```

```
long2 <- as.data.frame(cbind(long$Id,long$Cea_measure))
colnames(long2) <- c("Id","Cea_measure")

df_wide2 <- merge(df_wide,long2,by = "Id", all.y =T)
dim(df_wide2)


df_long2.4$months <- df_long2.4$time_cea_month
df_long2.4 <- df_long2.4 %>% dplyr::select("Id","Cea_measure","Cea_measure_ln","months","AgeOK","Age_cat","Sex_cat","Smoke_imp",
              "Tumor_imp","Stadium_cat1","Stadium_cat2", "ResectionMargeFree_imp", "Stadium_two","Stadium_three","Stadium_four","Smoke_1","Smoke_2",
              "Smoke_imp_cat","CeaBaseline_imp_log","month_rec_max60", "cen_rec_1826")

head(df_long2.4)

# check which variables are factors
(l <- sapply(df_long2.4, function(x) is.factor(x)))

##########################################################################################################
# Creating the training data set (80 %) and testing data set (20 %)
##########################################################################################################

#----------------------------------------------------------------
set.seed(678867)

#Create training set (wide format)
train_wide <- df_wide2 %>%
  dplyr::sample_frac(.80)

#Create test set ((wide format))
test_wide  <- anti_join(df_wide2, train_wide, by = ’Id’)

#----------------------------------------------------------------
##Create training set (Long format)
train_long  <- anti_join(df_long2.4, test_wide, by = ’Id’)

##Create test set (Long format)
test_long  <- anti_join(df_long2.4, train_wide, by = ’Id’)

#----------------------------------------------------------------

# Sorting the data sets

train_wide <-  train_wide %>% dplyr::arrange(Id)
test_wide <-  test_wide %>% dplyr::arrange(Id)

train_long <-  train_long %>% dplyr::arrange(Id)
test_long <-  test_long %>% dplyr::arrange(Id)
#----------------------------------------------------------------

#######################################################################
#                 Data exploration
#######################################################################
#-------------------------------------------------------------------------
#         Descriptive statistics
#------------------------------------------------------------------------

summary(df_wide2[,c("time_rec","time_DFS","time_surv")])
summary(df_wide2$AgeOK)
summary(df_wide2$CeaBaseline1)
```

```
tab1(df_wide2$Age_cat, graph = F)
tab1(df_wide2$Sex_cat, graph = F)
tab1(df_wide2$Smoke_cat, graph = F)
tab1(df_wide2$Palliative_cat, graph = F)
tab1(df_wide2$Tumor_cat, graph = F)
tab1(df_wide2$Stadium_cat1, graph = F)


#******************************************************
#tumor Reccurence
#******************************************************
tabyl(df_wide2$cen_rec_1826, sort = TRUE)

by(df_wide2$AgeOK,df_wide2$cen_rec_1826, summary)
by(df_wide2$CeaBaseline1,df_wide2$cen_rec_1826, summary)


df_wide2 %>% tabyl(Sex_cat, cen_rec_1826)
df_wide2 %>% tabyl(Smoke_cat, cen_rec_1826)
df_wide2 %>% tabyl(Tumor_cat, cen_rec_1826)
df_wide2 %>% tabyl(Stadium_cat1, cen_rec_1826)
df_wide2 %>% tabyl(ResectionMargeFree_cat, cen_rec_1826)

tabpct(df_wide2$Age_cat, df_wide2$cen_rec_1826, percent = "col", graph = F)
tabpct(df_wide2$Sex_cat, df_wide2$cen_rec_1826, percent = "col", graph = F)
tabpct(df_wide2$Smoke_cat, df_wide2$cen_rec_1826, percent = "col", graph = F)
tabpct(df_wide2$Tumor_cat, df_wide2$cen_rec_1826, percent = "col", graph = F)
tabpct(df_wide2$Stadium_cat1, df_wide2$cen_rec_1826, percent = "col", graph = F)
tabpct(df_wide2$ResectionMargeFree_cat, df_wide2$cen_rec_1826, percent = "col", graph = F)


#******************************************************
#Death event
#******************************************************

tabyl(df_wide2$cen_surv_1826, sort = TRUE)

by(df_wide2$AgeOK,df_wide2$cen_surv_1826, summary)
by(df_wide2$CeaBaseline1,df_wide2$cen_surv_1826, summary)

df_wide2 %>% tabyl(Sex_cat, cen_surv_1826)
df_wide2 %>% tabyl(Smoke_cat, cen_surv_1826)
df_wide2 %>% tabyl(Tumor_cat, cen_surv_1826)
df_wide2 %>% tabyl(Stadium_cat1, cen_surv_1826)
df_wide2 %>% tabyl(ResectionMargeFree_cat, cen_surv_1826)

tabpct(df_wide2$Age_cat, df_wide2$cen_surv_1826, percent = "col", graph = F)
tabpct(df_wide2$Sex_cat, df_wide2$cen_surv_1826, percent = "col", graph = F)
tabpct(df_wide2$Smoke_cat, df_wide2$cen_surv_1826, percent = "col", graph = F)
tabpct(df_wide2$Tumor_cat, df_wide2$cen_surv_1826, percent = "col", graph = F)
tabpct(df_wide2$Stadium_cat1, df_wide2$cen_surv_1826, percent = "col", graph = F)
tabpct(df_wide2$ResectionMargeFree_cat, df_wide2$cen_surv_1826, percent = "col", graph = F)


#******************************************************
#Disease Free Survival (event is death or reccurence))
#******************************************************

tabyl(df_wide2$cen_dfs_1826, sort = TRUE)
```

```
by(df_wide2$AgeOK,df_wide2$cen_dfs_1826, summary)
by(df_wide2$CeaBaseline1,df_wide2$cen_dfs_1826, summary)

df_wide2 %>% tabyl(Sex_cat, cen_dfs_1826)
df_wide2 %>% tabyl(Smoke_cat, cen_dfs_1826)
df_wide2 %>% tabyl(Tumor_cat, cen_dfs_1826)
df_wide2 %>% tabyl(Stadium_cat1, cen_dfs_1826)
df_wide2 %>% tabyl(ResectionMargeFree_cat, cen_dfs_1826)

tabpct(df_wide2$Age_cat, df_wide2$cen_dfs_1826, percent = "col", graph = F)
tabpct(df_wide2$Sex_cat, df_wide2$cen_dfs_1826, percent = "col", graph = F)
tabpct(df_wide2$Smoke_cat, df_wide2$cen_dfs_1826, percent = "col", graph = F)
tabpct(df_wide2$Tumor_cat, df_wide2$cen_dfs_1826, percent = "col", graph = F)
tabpct(df_wide2$Stadium_cat1, df_wide2$cen_dfs_1826, percent = "col", graph = F)
tabpct(df_wide2$ResectionMargeFree_cat, df_wide2$cen_dfs_1826, percent = "col", graph = F)


#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
# The histogram of the distribuition of the Tumor recurrence event over time (5 years)

summary(df_wide2$month_rec_max60)
df_wide2$month_rec_max60_1 <- round(df_wide2$month_rec_max60)
summary(df_wide2$month_rec_max60_1)

p <- df_wide2 %>% filter(cen_rec_1826 == 1) %>% dplyr::select(month_rec_max60_1,cen_rec_1826)

#table(p$month_rec_max60_1);barplot(table(p$month_rec_max60_1))

dist_rec_event <- as.data.frame(table(p$month_rec_max60_1))

ggplot(data=p, aes(x=month_rec_max60_1)) +
  geom_histogram(breaks=seq(0, 60, by=1),
                 col="white",
                 fill="black",
                 alpha = .2) +
  labs(title="", x="months", y="count") +
  xlim(c(0,60)) +
  ylim(c(0,25)) +
  theme_classic(base_size = 12)
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
# CEA baseline
length(which(is.na(df_wide2$CeaBaseline1))) # 278
length(which(!is.na(df_wide2$CeaBaseline1))) # 1822

#summary(df_wide2$CeaBaseline1)

#df_wide2$CeaBaseline_cat <- as.factor(ifelse(df_wide2$CeaBaseline1 >=  2.660, "Yes","No"))
#tab1(df_wide2$CeaBaseline_cat)

##################################################################################################################
#                 K.M Plots                                                                                  #
##################################################################################################################
#--------------------------------------------------------------------------------------------------------------
#overall (Reccurence)
#--------------------------------------------------------------------------------------------------------------
rec.KM <- survfit(Surv(month_rec_max60, cen_rec_1826 ) ~ 1, data=df_wide2)
summary(rec.KM)
ggsurvplot(rec.KM,data=df_wide2,xlab="Survival time (Months)",legend.labs =c("Tumor recurrence"),legend.title="", ylab="Survival probability")
```

```
plot(rec.KM)
abline(v = c(12,24,36,48,60),lty = 3)
abline(h = c(.93,.86,.835,.815,.80),lty = 3, col = c(1:5))

#sex

rec_sex.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ Sex, data=df_wide2)
print(rec_sex.KM)

ggsurvplot(rec_sex.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Female","Male"),legend.title="Sex:",pval = TRUE)

#Smoking status
rec_smoke.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ Smoke_imp, data=df_wide2)
print(rec_smoke.KM)

ggsurvplot(rec_smoke.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Non smoker","Current","Former"),legend.title="Smoke:",pval = TRUE)

#Tumor_cat
rec_tumor.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ Tumor_imp, data=df_wide2)
print(rec_tumor.KM)

ggsurvplot(rec_tumor.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Adenocarcinoma","Mucinous carcinoma"),
           legend.title="Tumor stage:",pval = TRUE)

#Tumor stage
rec_Stadium.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ as.factor(Stadium_cat1), data=df_wide2)
print(rec_Stadium.KM)
ggsurvplot(rec_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Stage 1","Stage 2","Stage 3","Stage 4"),legend.title="Cancer stage:",pval = TRUE)

#Resection Margin
rec_Stadium.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ ResectionMargeFree, data=df_wide2)
print(rec_Stadium.KM)
ggsurvplot(rec_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("No","Yes"),legend.title="Resection margin:",pval = TRUE)

#Pre-CEA measurement

rec_base.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ df_wide2$CeaBaseline_cat, data=df_wide2)
print(rec_base.KM)
ggsurvplot(rec_base.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("No","Yes"),legend.title="Baseline CEA:",pval = TRUE)

#Age_cat
rec_age.KM <- survfit(Surv(month_rec_max60, cen_rec_1826) ~ df_wide2$Age_cat, data=df_wide2)
print(rec_age.KM)
ggsurvplot(rec_age.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.title="Age:",pval = TRUE)

#-----------------------------------------------------------------------------------------------------------
#overall (Death)
#-----------------------------------------------------------------------------------------------------------
surv.KM <- survfit(Surv(month_surv_max60, cen_surv_1826 ) ~ 1, data=df_wide2)
summary(surv.KM)
ggsurvplot(surv.KM,data=df_wide2,xlab="Survival time (Months)",legend.labs =c("Tumor survurrence"),legend.title="", ylab="Survival probability")

#sex
```

```
surv_sex.KM <- survfit(Surv(month_surv_max60, cen_surv_1826) ~ Sex, data=df_wide2)
print(surv_sex.KM)

ggsurvplot(surv_sex.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Female","Male"),legend.title="Sex:",pval = TRUE)

#Smoking status
surv_smoke.KM <- survfit(Surv(month_surv_max60, cen_surv_1826) ~ Smoke_imp, data=df_wide2)
print(surv_smoke.KM)

ggsurvplot(surv_smoke.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Non smoker","Current","Former"),legend.title="Smoke:",pval = TRUE)

#Tumor type
surv_tumor.KM <- survfit(Surv(month_surv_max60, cen_surv_1826) ~ Tumor_imp, data=df_wide2)
print(surv_tumor.KM)

ggsurvplot(surv_tumor.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Adenocarcinoma","Mucinous carcinoma"),
           legend.title="Tumor stage:",pval = TRUE)

#Tumor stage
surv_Stadium.KM <- survfit(Surv(month_surv_max60, cen_surv_1826) ~ as.factor(Stadium_cat1), data=df_wide2)
print(surv_Stadium.KM)
ggsurvplot(surv_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Stage 1","Stage 2","Stage 3","Stage 4"),legend.title="Cancer stage:",pval = TRUE)

#Resection Margin

surv_Stadium.KM <- survfit(Surv(month_surv_max60, cen_surv_1826) ~ ResectionMargeFree_imp, data=df_wide2)
print(surv_Stadium.KM)
ggsurvplot(surv_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("No","Yes"),legend.title="Resection margin:",pval = TRUE)

#-----------------------------------------------------------------------------------------------------------
# Disease Free Survival (event is death or reccurence))
#-----------------------------------------------------------------------------------------------------------
dfs.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ 1, data=df_wide2)
print(dfs.KM)
ggsurvplot(dfs.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability")


#Sex
dfs_sex.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ Sex, data=df_wide2)
print(dfs_sex.KM)

ggsurvplot(dfs_sex.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Female","Male"),legend.title="Sex:", pval = TRUE)

#Smoking status
dfs_smoke.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ Smoke_imp, data=df_wide2)
print(dfs_smoke.KM)

ggsurvplot(dfs_smoke.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Non smoker","Current","Former"),legend.title="Smoke:",pval = TRUE)

#Tumor type
dfs_tumor.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ Tumor_imp, data=df_wide2)
print(dfs_tumor.KM)

ggsurvplot(dfs_tumor.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Adenocarcinoma","Mucinous carcinoma"),
           legend.title="Tumor Stage:",pval = TRUE)

#Tumor stage
dfs_Stadium.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ as.factor(Stadium_cat1), data=df_wide2)
print(dfs_Stadium.KM)
ggsurvplot(dfs_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("Stage 1","Stage 2","Stage 3","Stage 4"),legend.title="Cancer stage:",pval = TRUE)

#Resection Margin
dfs_Stadium.KM <- survfit(Surv(month_DFS_max60, cen_dfs_1826) ~ ResectionMargeFree_imp, data=df_wide2)
print(dfs_Stadium.KM)
ggsurvplot(dfs_Stadium.KM,data=df_wide2,xlab="Survival time (Months)", ylab="Survival probability",
           legend.labs =c("No","Yes"),legend.title="Resection margin:",pval = TRUE)

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
#######################################################################
#                  LDA Data exploration
#######################################################################

# number of visits
visit <- as.data.frame(table(df_long2.4$Id))


visit <-  visit %>%
  arrange(Freq)

summary(visit$Freq)

# Frequency of number of CEA measurement per patient
ggplot(data=visit, aes(x=Freq)) +
  geom_histogram(breaks=seq(0, 25, by=1),
                 col="white",
                 fill="black",
                 alpha = .2) +
  labs(title="", x="Number of measurements", y="Number of patients") +
  xlim(c(0,25)) +
  ylim(c(0,250)) +
  theme_classic(base_size = 12)

#----------------------------------------------------------------------------------
#******************* Individual profiles (after log transformation) *******************
#----------------------------------------------------------------------------------
#library(data.table)
set.seed(84948)
#-------------------------------------------------------------------------------------------
# select the x % of the data

plot_sample <- df_wide2 %>%
  dplyr::sample_frac(.98)
plot_sample1  <- anti_join(train_long, plot_sample, by = ’Id’)


xyplot(Cea_measure_ln ~ months, group = Id, data = plot_sample1,ylim = c(0,6),
       scales = list(tck = c(-1, 0)),
       panel = function(x, y, ...) {
         panel.xyplot(x, y, type = "l", col = "gray", ...)
         panel.loess(x, y, col = 2, lwd = 2)
       }, xlab = "Follow-up time (Months)", ylab = "ln(CEA+1) Ug/L")
#-------------------------------------------------------------------------------------------
```

```
#-------------------------------------------------------------------------------------------
#(Tumor reccurence) #

x.rec1 <- df_long2.4[which(df_long2.4$cen_rec_1826 == 1),]

set.seed(84948)

#Create training set (wide format)
x.rec1_sample <- df_wide2 %>%
  dplyr::sample_frac(.86)
x.rec1_sample1  <- anti_join(x.rec1, x.rec1_sample, by = ’Id’)


xyplot(Cea_measure_ln ~ months, group = Id, data = x.rec1_sample1,ylim = c(0,4.5),
       scales = list(tck = c(-1, 0),y=list(at=c(0.5,1,1.5,2,2.5,3,3.5,4,4.5,5))),
       panel = function(x, y, ...) {
         panel.xyplot(x, y, type = "l", lwd = 1,col = "gray", ...)
         panel.loess(x, y, col = 2, lwd = 2)
       },xlab = "Follow-up time (Months)", ylab = "ln(CEA+1) Ug/L")

#-------------------------------------------------------------------------------------------

# patients Without tumor recurrence

x.rec2 <- df_long2.4[which(df_long2.4$cen_rec_1826 == 0),]

set.seed(84948)

#Create training set (wide format)
x.rec2_sample <- df_wide2 %>%
  dplyr::sample_frac(.97)
x.rec2_sample1  <- anti_join(x.rec2, x.rec2_sample, by = ’Id’)


xyplot(Cea_measure_ln ~ months, group = Id, data = x.rec2_sample1,ylim = c(0,4.5),
       scales = list(tck = c(-1, 0),y=list(at=c(0.5,1,1.5,2,2.5,3,3.5,4,4.5))),
       panel = function(x, y, ...) {
         panel.xyplot(x, y, type = "l", lwd = 1,col = "gray", ...)
         panel.loess(x, y, col = 2, lwd = 2)
       },xlab = "Follow-up time (Months)", ylab = "ln(CEA+1) Ug/L")
#------------------------------------------------------------------------------------------
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

#---------------------------------------------
#---------------- LDA submodel ----------------
#---------------------------------------------

lmeFit1 <- lme(Cea_measure_ln  ~  Age_cat+Sex_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
               random = ~ 1 | Id, data = train_long)
summary(lmeFit1)


# Without age variables
lmeFit2 <- lme(Cea_measure_ln  ~ Sex_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
               random = ~ 1 | Id, data = train_long)
summary(lmeFit2)


# Without sex
lmeFit3 <- lme(Cea_measure_ln  ~ Age_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
               random = ~ 1 | Id, data = train_long)
```

```
summary(lmeFit3)


m1 <- update(lmeFit1, method = "ML") #(all variables)
m2 <- update(lmeFit2, method = "ML") #(without age variable)
m3 <- update(lmeFit3, method = "ML") #(Without sex variables)


1-pchisq(-2*(m2$logLik-m1$logLik),1) #(without age variable)
1-pchisq(-2*(m2$logLik-m1$logLik),1) #(Without sex variables)

#--------------------------------------------------------------------
#test for need random slope
#--------------------------------------------------------------------

lmeFit11 <- lme(Cea_measure_ln  ~ Age_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
                random = ~ 1 | Id, data = train_long)
summary(lmeFit11)

lmeFit12 <- lme(Cea_measure_ln  ~ Age_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
                random = ~ months | Id, data = train_long)
summary(lmeFit12)

lmeFit11$logLik
lmeFit12$logLik

# Mixture chi square
0.5*(1-pchisq(1990.364,1))+0.5*(1-pchisq(1990.364,2))

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

######################################################################################
#         Seperate  Survival submodel (Cox PH regression Model)                                  #
######################################################################################

#functional form of size; plots based on the martingale residuals

empty.PH <- coxph(Surv(month_rec_max60, cen_rec_1826)~1, data = train_wide)
mart.res <- resid(empty.PH)

plot(train_wide$CeaBaseline_imp,mart.res, xlab = "Pre-surgery CEA measurement", ylab = "Martingale residuals")
lines(lowess(train_wide$CeaBaseline_imp,mart.res,iter=0,f=0.6))

# log
plot(train_wide$CeaBaseline_imp_log,mart.res, xlab = "CEA baseline measurement", ylab = "Martingale residuals")
lines(lowess(train_wide$CeaBaseline_imp_log,mart.res,iter=0,f=0.6))


# add all the baseline characteristics to the null model
coxFit10 <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat +Sex_cat + Smoke_imp + as.factor(Tumor_imp) + Stadium_cat1
                  + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                  data = train_wide, x = TRUE,model = TRUE)
summary(coxFit10)


# Using the schoenfield residuals to  check the model diagnostic
gbcs.PHfit <- cox.zph(coxFit10, transform="log")
gbcs.PHfit

plot(gbcs.PHfit)
```

```
#overal fit of the model:
devres <- residuals(coxFit10,type="deviance")
fitval<- predict(coxFit10,type="lp")
plot(fitval,devres)

######################################################################################################
#                  Overall effect of the covariates in the Cox PH model
#######################################################################################################

coxFit10 <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat +Sex_cat + Smoke_imp + Tumor_imp + Stadium_cat1
                  + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                  data = train_wide, x = TRUE,model = TRUE)
summary(coxFit10)



# no including age
coxFit_noage <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Sex_cat + Smoke_imp + Tumor_imp + Stadium_cat1
                      + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                      data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_noage)

# no including smoke
coxFit_nosmoke <- coxph(Surv(month_rec_max60, cen_rec_1826) ~   Age_cat + Sex_cat +Tumor_imp + Stadium_cat1
                      + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                      data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_nosmoke)

# no including sex
coxFit_nosex <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat  + Smoke_imp + Tumor_imp + Stadium_cat1
                      + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                      data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_nosex)

# no including smoke
coxFit_nosmoke <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat + Sex_cat +  Tumor_imp + Stadium_cat1
                        + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                        data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_nosmoke)

# no including tumor type
coxFit_notumortype <- coxph(Surv(month_rec_max60, cen_rec_1826) ~   Age_cat + Sex_cat + Smoke_imp + Stadium_cat1
                           + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                           data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_notumortype)

# no including tumor stage
coxFit_notumorstage <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat +Sex_cat + Smoke_imp +Tumor_imp
                             + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                             data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_notumorstage)

# no including pre-surgery
coxFit_noprecea <- coxph(Surv(month_rec_max60, cen_rec_1826) ~  Age_cat +Sex_cat + Smoke_imp + Tumor_imp + Stadium_cat1
                         + strata(as.factor(ResectionMargeFree_imp)),
                         data = train_wide, x = TRUE,model = TRUE)
summary(coxFit_noprecea)

# Using the log likelihood ratio
-2*(coxFit_noage$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_noage$loglik[2]-coxFit10$loglik[2]),1)

-2*(coxFit_nosex$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_nosex$loglik[2]-coxFit10$loglik[2]),1)
```

```
-2*(coxFit_nosmoke$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_nosmoke$loglik[2]-coxFit10$loglik[2]),2)

-2*(coxFit_notumortype$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_notumortype$loglik[2]-coxFit10$loglik[2]),1)

-2*(coxFit_notumorstage$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_notumorstage$loglik[2]-coxFit10$loglik[2]),3)

-2*(coxFit_noprecea$loglik[2]-coxFit10$loglik[2]);1-pchisq(-2*(coxFit_noprecea$loglik[2]-coxFit10$loglik[2]),1)


#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

# Concordance Index
library(dynpred)
library(prodlim)
library(pec)

cox1 <- coxph(Surv(month_rec_max60, cen_rec_1826)~ Age_cat +Sex_cat + Smoke_imp + as.factor(Tumor_imp) + Stadium_cat1
              + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log, x=TRUE, data = train_wide)

A1  <- pec::cindex(cox1,
                   formula=Surv(month_rec_max60, cen_rec_1826)~ Age_cat +Sex_cat + Smoke_imp + as.factor(Tumor_imp) + Stadium_cat1
                   + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                   data=test_wide,
                   eval.times=36)
A1

#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>


########################################################################################
#                                 JOINT MODELLING                                      #
########################################################################################


#            Joint model with no baseline covariates
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

lmeFit1 <- lme(Cea_measure_ln  ~ months,
               random = ~ months | Id, data = train_long)
summary(lmeFit1)
print.lmm.wald(lmeFit1)

#train_wide$month_rec_max60 <- round(train_wide$month_rec_max60)
coxFit1 <- coxph(Surv(month_rec_max60, cen_rec_1826) ~ 1,
                 data = train_wide, x = TRUE,model = TRUE)
summary(coxFit1)
#print.HRCIs(coxFit1)

#---------------------------
jointFit.1 <- jointModel(lmeFit1, coxFit1,
                         timeVar = "months", method = "weibull-PH-aGH")
summary(jointFit.1)
print.joint.lda(jointFit.1)
print.joint.event(jointFit.1)

plot(jointFit.1)
#-----------------------------


#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
```

```
#------------- Final Joint Models assuming  Weibull assumption in baseline hazard ----------------------
#>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>



lmeFit <- lme(Cea_measure_ln  ~ Age_cat+Smoke_imp+Stadium_cat1+months+ months*Age_cat,
              random = ~ months| Id, data = train_long)
summary(lmeFit)
anova.lme(lmeFit, type = "marginal", adjustSigma = F)
print.lmm.wald(lmeFit)

coxFit <- coxph(Surv(month_rec_max60, cen_rec_1826) ~ Age_cat + Sex_cat + Smoke_imp + Tumor_imp+Stadium_cat1
                + strata(as.factor(ResectionMargeFree_imp))+ CeaBaseline_imp_log,
                data = train_wide, x = TRUE,model = TRUE)
summary(coxFit)
print.HRCIs(coxFit)

#--------------------------------------------------------------------------
#                     Weibull assumption weibull-PH-aGH;
#--------------------------------------------------------------------------



jointFit1 <- jointModel(lmeFit, coxFit,
                        timeVar = "months", method = "weibull-PH-aGH")
summary(jointFit1)
print.joint.lda(jointFit1)
print.joint.event(jointFit1)

################### True slope value trajectory association ##############################
dForm1 <- list(fixed = ~ Age_cat, random = ~1, indFixed = c(8:9), indRandom = 2)
#
jointFit1.1 <- update(jointFit1, param = "slope", derivForm = dForm1)
summary(jointFit1.1)
print.joint.lda(jointFit1.1)
print.joint.event(jointFit1.1)

################### True value plus the true slope value trajectory association ##############################
jointFit1.2 <- update(jointFit1, param = "both", derivForm = dForm1)
summary(jointFit1.2)
print.joint.lda(jointFit1.2)
print.joint.event(jointFit1.2)

plot(jointFit1.2)

######
BIC(jointFit1,jointFit1.1,jointFit1.2);AIC(jointFit1,jointFit1.1,jointFit1.2)




#-----------------------------------------------------------------------------
# Overall effects of covariates in current plus slope joint model formulation
#-----------------------------------------------------------------------------

#coxFit_noage coxFit_nosmoke coxFit_nosex coxFit_notumortype coxFit_notumorstage coxFit_noprecea

jointFit1.2_noage <- update(jointFit1.2, survObject = coxFit_noage,
                        parameterization = "both", derivForm = dForm1)

jointFit1.2_nosmoke<- update(jointFit1.2, survObject = coxFit_nosmoke,
                            parameterization = "both", derivForm = dForm1)
```

```
jointFit1.2_nosex <- update(jointFit1.2, survObject = coxFit_nosex,
                             parameterization = "both", derivForm = dForm1)

jointFit1.2_notumortype <- update(jointFit1.2, survObject = coxFit_notumortype,
                            parameterization = "both", derivForm = dForm1)

jointFit1.2_notumorstage <- update(jointFit1.2, survObject = coxFit_notumorstage,
                                  parameterization = "both", derivForm = dForm1)

jointFit1.2_noprecea <- update(jointFit1.2, survObject = coxFit_noprecea,
                                   parameterization = "both", derivForm = dForm1)


anova(jointFit1.2_noage,jointFit1.2) # overall effect of the age
anova(jointFit1.2_nosmoke,jointFit1.2) # overall effect of the smoking status
anova(jointFit1.2_nosex,jointFit1.2) # overall effect of the sex
anova(jointFit1.2_notumortype,jointFit1.2) # overall effect of the tumor type
anova(jointFit1.2_notumorstage,jointFit1.2) # overall effect of the tumor stage
anova(jointFit1.2_noprecea,jointFit1.2) # overall effect of the pre-suregery
anova(jointFit1,jointFit1.2) # overall effect of the current value
anova(jointFit1.1,jointFit1.2) # overall effect of the slope

#---------------------------------------------------------------------------------------------------------------------------------------
# Joint Model discrimination accuarcy using the Area under the Curve. (Current value without baseline covariates)
#---------------------------------------------------------------------------------------------------------------------------------------

auc.roc.fit.1_7.3 <-  aucJM(jointFit.1, newdata = test_long, Tstart = 7, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 12 months
auc.roc.fit.1_17.3 <- aucJM(jointFit.1, newdata = test_long, Tstart = 17, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 24 months
auc.roc.fit.1_27.3 <- aucJM(jointFit.1, newdata = test_long, Tstart = 27, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 36 months
auc.roc.fit.1_37.3 <- aucJM(jointFit.1, newdata = test_long, Tstart = 37, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 45 months
auc.roc.fit.1_47.3 <- aucJM(jointFit.1, newdata = test_long, Tstart = 47, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 57 months
auc.roc.fit.1_57.3 <- aucJM(jointFit.1, newdata = test_long, Tstart = 57, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 57 months

months_t.1.3 <- c(10,20,30,40,50,60)
auc_t.1.3 <- c(auc.roc.fit.1_7.3$auc,auc.roc.fit.1_17.3$auc,auc.roc.fit.1_27.3$auc,
               auc.roc.fit.1_37.3$auc,auc.roc.fit.1_47.3$auc,auc.roc.fit.1_57.3$auc)
at_risk.1.3 <- c(auc.roc.fit.1_7.3$nr,auc.roc.fit.1_17.3$nr,auc.roc.fit.1_27.3$nr,
                 auc.roc.fit.1_37.3$nr,auc.roc.fit.1_47.3$nr,auc.roc.fit.1_57.3$nr)

auc.roc.fit.1.3 <- cbind(months_t.1.3,at_risk.1.3,auc_t.1.3)
auc.roc.fit.1.3 <- as.data.frame(auc.roc.fit.1.3)
colnames(auc.roc.fit.1.3) <- c("time","at_risk","AUC")

auc.roc.fit.1.3

ddi.1.3 <- dynCJM(jointFit.1, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)

#---------------------------------------------------------------------------
auc.roc.fit.1_4.6 <-  aucJM(jointFit.1, newdata = test_long, Tstart = 4, Dt =6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 12 months
auc.roc.fit.1_14.6 <- aucJM(jointFit.1, newdata = test_long, Tstart = 14, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 24 months
auc.roc.fit.1_24.6 <- aucJM(jointFit.1, newdata = test_long, Tstart = 24, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 36 months
auc.roc.fit.1_34.6 <- aucJM(jointFit.1, newdata = test_long, Tstart = 34, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 45 months
auc.roc.fit.1_44.6 <- aucJM(jointFit.1, newdata = test_long, Tstart = 44, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 57 months
auc.roc.fit.1_54.6 <- aucJM(jointFit.1, newdata = test_long, Tstart = 54, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 57 months


months_t.1.6 <- c(10,20,30,40,50,60)
auc_t.1.6 <- c(auc.roc.fit.1_4.6$auc,auc.roc.fit.1_14.6$auc,auc.roc.fit.1_24.6$auc,
               auc.roc.fit.1_34.6$auc,auc.roc.fit.1_44.6$auc,auc.roc.fit.1_54.6$auc)
at_risk.1.6 <- c(auc.roc.fit.1_4.6$nr,auc.roc.fit.1_14.6$nr,auc.roc.fit.1_24.6$nr,
                 auc.roc.fit.1_34.6$nr,auc.roc.fit.1_44.6$nr,auc.roc.fit.1_54.6$nr)
auc.roc.fit.1.6 <- cbind(months_t.1.6,at_risk.1.6,auc_t.1.6)
auc.roc.fit.1.6 <- as.data.frame(auc.roc.fit.1.6)
colnames(auc.roc.fit.1.6) <- c("time","at_risk","AUC")

auc.roc.fit.1.6

ddi.1.6 <- dynCJM(jointFit.1, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)

#--------------------------------------------------------------------------------------------------------------------------------------
# Model accuarcy using the Area under the Curve and DDI. (Current value log CEA value association
#-------------------------------------------------------------------------------------------------------------------------------------
auc.roc.fit_7.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 7, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit_17.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 17, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit_27.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 27, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit_37.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 37, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit_47.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 47, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit_57.3 <- aucJM(jointFit1, newdata = test_long, Tstart = 57, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months

months_t.3 <- c(10,20,30,40,50,60)
auc_t.3 <- c(auc.roc.fit_7.3$auc,auc.roc.fit_17.3$auc,auc.roc.fit_27.3$auc,
             auc.roc.fit_37.3$auc,auc.roc.fit_47.3$auc,auc.roc.fit_57.3$auc)
at_risk.3 <- c(auc.roc.fit_7.3$nr,auc.roc.fit_17.3$nr,auc.roc.fit_27.3$nr,
               auc.roc.fit_37.3$nr,auc.roc.fit_47.3$nr,auc.roc.fit_57.3$nr)

auc.roc.fit.3 <- cbind(months_t.3,at_risk.3,auc_t.3)
auc.roc.fit.3 <- as.data.frame(auc.roc.fit.3)
colnames(auc.roc.fit.3) <- c("time","at_risk","AUC")

auc.roc.fit.3

ddi.3 <- dynCJM(jointFit1, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)

#----------------------------------------------------------------------------
auc.roc.fit_4.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 4, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit_14.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 14, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit_24.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 24, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit_34.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 34, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit_44.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 44, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit_54.6 <- aucJM(jointFit1, newdata = test_long, Tstart = 54, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months


months_t.6 <- c(10,20,30,40,50,60)
auc_t.6 <- c(auc.roc.fit_4.6$auc,auc.roc.fit_14.6$auc,auc.roc.fit_24.6$auc,
             auc.roc.fit_34.6$auc,auc.roc.fit_44.6$auc,auc.roc.fit_54.6$auc)
at_risk.6 <- c(auc.roc.fit_4.6$nr,auc.roc.fit_14.6$nr,auc.roc.fit_24.6$nr,
               auc.roc.fit_34.6$nr,auc.roc.fit_44.6$nr,auc.roc.fit_44.6$nr)

auc.roc.fit.6 <- cbind(months_t.6,at_risk.6,auc_t.6)
auc.roc.fit.6 <- as.data.frame(auc.roc.fit.6)
colnames(auc.roc.fit.6) <- c("time","at_risk","AUC")

auc.roc.fit.6

ddi.6 <- dynCJM(jointFit1, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)
#------------------------------------------------------------------------------------------------------------------------------------------------


# Model accuarcy using the Area under the Curve. (slope value)
#------------------------------------------------------------------------------------------------------------------------
```

```
auc.roc.fit1_7.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 7, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit1_17.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 17, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit1_27.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 27, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit1_37.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 37, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit1_47.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 47, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit1_57.3 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 57, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months


months_t1.3 <- c(10,20,30,40,50,60)
auc_t1.3 <- c(auc.roc.fit1_7.3$auc,auc.roc.fit1_17.3$auc,auc.roc.fit1_27.3$auc,
              auc.roc.fit1_37.3$auc,auc.roc.fit1_47.3$auc,auc.roc.fit1_57.3$auc)
at_risk1.3 <- c(auc.roc.fit1_7.3$nr,auc.roc.fit1_17.3$nr,auc.roc.fit1_27.3$nr,
                auc.roc.fit1_37.3$nr,auc.roc.fit1_47.3$nr,auc.roc.fit1_57.3$nr)

auc.roc.fit1.3 <- cbind(months_t1.3,at_risk1.3,auc_t1.3)
auc.roc.fit1.3 <- as.data.frame(auc.roc.fit1.3)
colnames(auc.roc.fit1.3) <- c("time","at_risk","AUC")

auc.roc.fit1.3

ddi1.3 <- dynCJM(jointFit1.1, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi1.3

#---------------------------------------------------------------------------
auc.roc.fit1_4.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 4, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit1_14.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 14, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit1_24.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 24, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit1_34.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 34, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit1_44.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 44, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit1_54.6 <- aucJM(jointFit1.1, newdata = test_long, Tstart = 54, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months


months_t1.6 <- c(10,20,30,40,50,60)
auc_t1.6 <- c(auc.roc.fit1_4.6$auc,auc.roc.fit1_14.6$auc,auc.roc.fit1_24.6$auc,
              auc.roc.fit1_34.6$auc,auc.roc.fit1_44.6$auc,auc.roc.fit1_54.6$auc)
at_risk1.6 <- c(auc.roc.fit1_4.6$nr,auc.roc.fit1_14.6$nr,auc.roc.fit1_24.6$nr,
                auc.roc.fit1_34.6$nr,auc.roc.fit1_44.6$nr,auc.roc.fit1_44.6$nr)

auc.roc.fit1.6 <- cbind(months_t1.6,at_risk1.6,auc_t1.6)
auc.roc.fit1.6 <- as.data.frame(auc.roc.fit1.6)
colnames(auc.roc.fit1.6) <- c("time","at_risk","AUC")

auc.roc.fit1.6

ddi1.6 <- dynCJM(jointFit1.1, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)
ddi1.6

# Model accuarcy using the Area under the Curve. (Current value plus slope)
#-----------------------------------------------------------------------------------------------------------------------

auc.roc.fit2_7.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 7, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit2_17.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 17, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit2_27.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 27, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit2_37.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 37, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit2_47.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 47, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit2_57.3 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 57, Dt = 3,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months


months_t2.3 <- c(10,20,30,40,50,60)
auc_t2.3 <- c(auc.roc.fit2_7.3$auc,auc.roc.fit2_17.3$auc,auc.roc.fit2_27.3$auc,
              auc.roc.fit2_37.3$auc,auc.roc.fit2_47.3$auc,auc.roc.fit2_57.3$auc)
```

```
at_risk2.3 <- c(auc.roc.fit2_7.3$nr,auc.roc.fit2_17.3$nr,auc.roc.fit2_27.3$nr,
                auc.roc.fit2_37.3$nr,auc.roc.fit2_47.3$nr,auc.roc.fit2_57.3$nr)

auc.roc.fit2.3 <- cbind(months_t2.3,at_risk2.3,auc_t2.3)
auc.roc.fit2.3 <- as.data.frame(auc.roc.fit2.3)
colnames(auc.roc.fit2.3) <- c("time","at_risk","AUC")

auc.roc.fit2.3

ddi2.3 <- dynCJM(jointFit1.2, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi2.3

#---------------------------------------------------------------------------

auc.roc.fit2_4.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 4, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 10 months
auc.roc.fit2_14.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 14, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 20 months
auc.roc.fit2_24.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 24, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 30 months
auc.roc.fit2_34.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 34, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 40 months
auc.roc.fit2_44.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 44, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 50 months
auc.roc.fit2_54.6 <- aucJM(jointFit1.2, newdata = test_long, Tstart = 54, Dt = 6,idVar = "Id",simulate = TRUE, M = 200) # AUC at 60 months


months_t2.6 <- c(10,20,30,40,50,60)
auc_t2.6 <- c(auc.roc.fit2_4.6$auc,auc.roc.fit2_14.6$auc,auc.roc.fit2_24.6$auc,
              auc.roc.fit2_34.6$auc,auc.roc.fit2_44.6$auc,auc.roc.fit2_54.6$auc)
at_risk2.6 <- c(auc.roc.fit2_4.6$nr,auc.roc.fit2_14.6$nr,auc.roc.fit2_24.6$nr,
                auc.roc.fit2_34.6$nr,auc.roc.fit2_44.6$nr,auc.roc.fit2_44.6$nr)

auc.roc.fit2.6 <- cbind(months_t2.6,at_risk2.6,auc_t2.6)
auc.roc.fit2.6 <- as.data.frame(auc.roc.fit2.6)
colnames(auc.roc.fit2.6) <- c("time","at_risk","AUC")

auc.roc.fit2.6

ddi2.6 <- dynCJM(jointFit1.2, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)
ddi2.6

#-----------------------------------------------------------------------------------
#  Dynamic prediction (Subject 16  and subject 1022)
#------------------------------------------------------------------------------------

#-------------> Conditional survival probabilities <----------
       #VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV#

----------------------------------------------------------------
#  Information up to 20 months
#----------------------------------------------------------------
par(mfrow = c(1,2))
fit1520.1 <- survfitJM(jointFit1.2, newdata = test_long[which(test_long$Id == 16 & test_long$months <= 20),], idVar = "Id")
plot(fit1520.1, estimator = "mean", include.y = TRUE,
     xlab = "Time (months)",
     conf.int = TRUE, fill.area = TRUE, col.area = "lightgrey")
mtext("ln(CEA+1) Ug/L",side=2,line=2)

fit1755.1 <- survfitJM(jointFit1.2, newdata = test_long[which(test_long$Id == 1022  & test_long$months <= 20),], idVar = "Id")
plot(fit1755.1, estimator = "mean", include.y = TRUE,
     xlab = "Time (months)",
     conf.int = TRUE, fill.area = TRUE, col.area = "lightgrey")
mtext("ln(CEA+1) Ug/L",side=2,line=2)

#------------------------------------------------------------------------------------
```

```
#  Information up to 40 months
#------------------------------------------------------------------------------------
fit1520.2 <- survfitJM(jointFit1.2, newdata = test_long[which(test_long$Id == 16 & test_long$months <= 40),], idVar = "Id")
plot(fit1520.2, estimator = "mean", include.y = TRUE,
     xlab = "Time (months)",
     conf.int = TRUE, fill.area = TRUE, col.area = "lightgrey")
mtext("ln(CEA+1) Ug/L",side=2,col="black",line=2)

fit1755.2 <- survfitJM(jointFit1.2, newdata = test_long[which(test_long$Id == 1022 & test_long$months <= 40),], idVar = "Id")
plot(fit1755.2, estimator = "mean", include.y = TRUE,
     xlab = "Time (months)",
     conf.int = TRUE, fill.area = TRUE, col.area = "lightgrey")
mtext("ln(CEA+1) Ug/L",side=2,col="black",line=2)

#-------------> longitudinal CEA measurements predictions <----------
           #VVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVVV#

#----------------------------------------------------------------------------------------
#par(mfrow = c(2,2))

lfit1 <- predict(jointFit1.2, newdata = test_long[which(test_long$Id == 16 & test_long$months <= 20),],
                 type = "Subject", interval = "conf", returnData = TRUE,idVar = "Id")

last.time <- with(lfit1, months[!is.na(low)][1])
p1 <- xyplot(pred + low + upp ~ months, data = lfit1, type = "l",xlab = "Time (Months)",ylim = c(0,3),
             ylab = "Predicted log(CEA)", main = "Subject 16",
             lty = c(1,2,2), col = c(2,1,1), lwd = 2,abline = list(v = last.time, lty = 3))

lfit1.1 <- predict(jointFit1.2, newdata = test_long[which(test_long$Id == 1022 & test_long$months <= 20),],
                   type = "Subject", interval = "conf", returnData = TRUE,idVar = "Id")

last.time <- with(lfit1.1, months[!is.na(low)][1])
p2 <- xyplot(pred + low + upp ~ months, data = lfit1.1, type = "l",xlab = "Time (Months)",ylim = c(0,3),
             ylab = "Predicted log(CEA)", main = "Subject 1022",
             lty = c(1,2,2), col = c(2,1,1), lwd = 2,abline = list(v = last.time, lty = 3))

#------------------------------------------------------------------------------------------------------------------------
lfit2 <- predict(jointFit1.2, newdata = test_long[which(test_long$Id == 16 & test_long$months <= 40),],
                 type = "Subject", interval = "conf", returnData = TRUE,idVar = "Id")

last.time <- with(lfit2, months[!is.na(low)][1])
p3 <- xyplot(pred + low + upp ~ months, data = lfit2, type = "l",xlab = "Time (Months)",ylim = c(0,3),
             ylab = "Predicted log(CEA)",main = "Subject 16",
             lty = c(1,2,2), col = c(2,1,1), lwd = 2,abline = list(v = last.time, lty = 3))

lfit2.1 <- predict(jointFit1.2, newdata = test_long[which(test_long$Id == 1022 & test_long$months <= 40),],
                   type = "Subject", interval = "conf", returnData = TRUE,idVar = "Id")

last.time <- with(lfit2.1, months[!is.na(low)][1])
p4 <-  xyplot(pred + low + upp ~ months, data = lfit2.1, type = "l",xlab = "Time (Months)",ylim = c(0,3),
              ylab = "Predicted log(CEA)",main = "Subject 1022",
              lty = c(1,2,2), col = c(2,1,1), lwd = 2,abline = list(v = last.time, lty = 3))


library(magrittr)
library(multipanelfigure)
figure1 <- multi_panel_figure(columns = 2, rows = 1, panel_label_type = "none",width = "auto",height = "auto")

figure1 %<>%
  fill_panel(p1, column = 1, row = 1) %<>%
  fill_panel(p2, column = 2, row = 1)
```

```
figure1

figure2 <- multi_panel_figure(columns = 2, rows = 1, panel_label_type = "none",width = "auto",height = "auto")

figure2 %<>%
  fill_panel(p3, column = 1, row = 1) %<>%
  fill_panel(p4, column = 2, row = 1)

figure2

#-----------------------------------------------------------------------------------------
#     Model diagnostic (Survival submodel) : current log CEA plus slope association
#-----------------------------------------------------------------------------------------
# Cox snell residuals

res <- residuals(jointFit1.2, process = "Event", type = "CoxSnell")
sfit <- survfit(Surv(res,cen_rec_1826)~1, data = train_wide)

plot(sfit, mark.time = FALSE, conf.int = TRUE, xlab = "Cox snell residuals", ylab = "Survival probablity",
     main = "", xlim=c(0,5))
curve(exp(-x), from = 0, to = max(train_wide$month_rec_max60), add = TRUE, col = "gray", lwd=2)

# Martingaele residuals


martRes<- residuals(jointFit1.2, process = "Event")
mi.t <- fitted(jointFit1.2, process = "Longitudinal",
               type = "EventTime")

plotResid <- function (x, y, col.loess = "black", ...) {
  plot(x, y, ...)
  lines(lowess(x, y), col = col.loess, lwd = 2)
  abline(h = 0, lty = 3, col = "grey", lwd = 2)
}


plotResid(mi.t, martRes, col.loess = "gray",
          ylab = "Martingale Residuals",
          xlab = "Subject-Specific Fitted Values Longitudinal Outcome")

xyplot(martRes ~ mi.t, type = c("p", "smooth"),col.loess = "gray",
       col = "black", lwd = 3, ylab = "Martingale Residuals",
       xlab = "Subject-Specific Fitted Values Longitudinal Outcome")

#---------------------------------------------------------------------------------------
#          Sentivity analysis                                   #
#---------------------------------------------------------------------------------------
#------------------------- > Baseline hazard (Weibull)

# -------  lagged effect --------------------------

jointFit1_lagged <- update(jointFit1, lag =1)
summary(jointFit1_lagged)

ddi.l1.3 <- dynCJM(jointFit1_lagged, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi.l1.3

ddi.l1.6 <- dynCJM(jointFit1_lagged, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)
ddi.l1.6
```

```
#


#--------------------------------------------------------------------------
#                     B-spline baseline hazard assumption
#----------------------------------------------------------------------

#################### Current value association ##############################
jointFit4 <- jointModel(lmeFit, coxFit,
                        timeVar = "months", method = "spline-PH-GH")
summary(jointFit4)


#################### Current slope value trajectory association ##############################
dFormI <- list(fixed = ~ Age_cat, random = ~1, indFixed = c(8:9), indRandom = 2)
#
jointFit4.1 <- update(jointFit4, param = "slope", derivForm = dForm1)
summary(jointFit4.1)


#################### Current value plus the Current slope value trajectory association ##############################
jointFit4.2 <- update(jointFit4, param = "both", derivForm = dForm1)
summary(jointFit4.2)

#################### lagged association ##############################
jointFit4_lagged <- update(jointFit4, lag =1)
summary(jointFit4_lagged)

#DDI

#
ddi.4.3 <- dynCJM(jointFit4, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi.4.6 <- dynCJM(jointFit4, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)

#
ddi.41.3 <- dynCJM(jointFit4.1, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi.41.6 <- dynCJM(jointFit4.1, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)

#
ddi.42.3 <- dynCJM(jointFit4.2, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi.42.6 <- dynCJM(jointFit4.2, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)

#
ddi.4l.3 <- dynCJM(jointFit4_lagged, newdata =  test_long, Dt = 3, t.max = 36,idVar = "Id",  M = 200)
ddi.4l.3
ddi.4l.6 <- dynCJM(jointFit4_lagged, newdata =  test_long, Dt = 6, t.max = 36,idVar = "Id",  M = 200)
ddi.4l.6

#-------------------------------------------------------------------------------------------------------
```