Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Flexible survival regression modelling: An application to lung cancer data

Relatore: Prof.ssa Giuliana Cortese
Dipartimento di Scienze Statistiche

Laureando: Marco Lazzarini
Matricola: 1222897

Anno accademico: 2022/2023

Contents

Introduction

Cox semi-parametric proportional hazards regression model represents one of the most widely used tools in the analysis of survival data. The underlying theory of profile likelihood allows one to effectively perform, given a certain set of regressors, a relatively easy-to-interpret estimation of the risk and survival curves and predictions for the outcome of interest.

Although such a model has been, and still is, much appreciated precisely because of its ease of use and interpretation, it is also acknowledged that said ease of use comes at the cost of some very stringent assumptions, which in fact pose uncomfortable limitations when it comes to having to face various real-world situations.

An important caveat to this theory is that the values of the regressors must be determined at time t_0, when the patient enters the study, and remain constant thereafter. However, there are numerous situations in which the effects of the variables included in the model are subject to time-dependence, thus resulting in the violation of the cardinal assumption of proportionality of the risks and consequently losing the capability of producing reliable predictions.

To accomodate covariates which may change their effects on the risk over time, special adjustments have to be done on the structure of the initial model. Furthermore, considering a whole class of non-multiplicative hazards models for the risk function can be a valuable option. On a more general level, this approach, in which some of the assumptions of a regression model are relaxed, is commonly referred to, in the statistics literature, as *flexible regression*.

In this dissertation, following some brief recalls on the main quantities and estimators used in survival analysis, the assumptions, diagnostics and limitations of the Cox proportional hazards regression model are discussed, along with the issue of handling time-dependent effects and exploring the main methods for evaluating and selecting the most appropriate functional form of the variables. Next, we show the usage of P-splines in the Cox model which allow a more dynamic modelling of nonlinear effects over time, followed by two examples of more flexible model families: *additive hazards models* and *multiplicative-additive hazards models*. Said models represent an effective solution for the aforementioned scenario, for they allow data to be modeled even if the effects of the variables on the risk are subject to time-dependence.

Comparison among the models, in terms of fitting and performance, is shown upon application to a set of real data concerning 181 stage I-IIIb NSCLC patients treated with (chemo-)radiotherapy between March 2007 and September 2013, in which blood-biomarkers related to hypoxia, inflammation, immune response and tumour load were reported. All patients participated in the Biobank project (Clinical trials.gov identifier: NCT01936571) launched in 2003.

1 Multiplicative hazards model

In this section we provide a presentation of the main element of the class of multiplicative hazards models, namely the Cox semi-parametric proportional hazards model.

1.1 Brief recalls on fundamental quantities and estimators in survival analysis

In the following, we recall the main quantities that define the basis for any descriptive and inferential procedure concerning survival data. For a more specific discussion we refer the reader to Klein & Moeschberger[16].

Survival data are data whose principal interest is the waiting time with respect to the occurrence of a set of one or more events, which are designated as interest events.
Censoring is a mechanism peculiar to this type of data and can occur in various forms: right, left, interval. In addition, it can be dependent or independent. Survival data whose observations exhibit right-handed censoring are characterized by the fact that at least one of the subjects in the study does not experience the event of interest within the observation period.
Let us focus on right censoring and define two variables known as $X \geq 0$ *waiting time to event* and $C \geq 0$ *censoring time*. According to what was explained earlier, only one of the two is observed, then:

$$T = \min(C, X) \geq 0 \tag{1}$$

is called *survival time*.
If censoring time C is predetermined and has the same value for all subjects, it's defined as *simple type I* right-hand censoring; this generally coincides with the end of the study. If the censoring times happen at two or more predetermined time points, we speak of *progressive type I* censoring. If the units enter the study at predetermined different times from each other but the censoring time is the same for all, we speak of *generic type I* censoring.
In *simple type II*, the study continues until the r-th event is observed, with $r < n$ where

n is the number of subjects. In *generic type II*, the n statistical units continue to be observed until the occurrence of r_1 events. $n_1 - r_1$ units are then censored among the remaining $n - r_1$; thus, $n - n_1$ units remain in the study. Subsequent r_2 events are then observed, and among the remaining units more $n_2 - r_2$ are censored, and so on. In the *random type*, all censoring times are random; this typically occurs when withdrawal and lost-to-follow-up cases are present in the study.

The **survival function** is a function which indicates the probability that an individual will survive beyond a certain period of time, considering the information collected up to that point. In other words, the survival function indicates the probability that an individual did *not* experience the event of interest (e.g., death or a disease) during the observation period.
The survival function is denoted by the letter S and can be defined as:

$$S(t) = P(T > t) \tag{2}$$

In other words, the survival function $S(t)$ is defined as the probability that the survival time is greater than t. Obviously, $S(0) = 1$, and $\lim_{x \to 1} S(x) = 0$.
The survival function can also be expressed in terms of: $S(t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function of survival time T.

The **hazard function** is a function which indicates the instantaneous probability that an individual suffers the event of interest in a given instant of time, considering the information collected up to that moment. In other words, the risk function indicates the speed at which the event of interest occurs in a given instant of time. Such function is frequently indicated with the letter h and can be defined as:

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3}$$

and represents the ratio of the probability that the event of interest occurs between t and $t + \Delta t$ (given that the individual has already survived up to t) and the length of the time interval Δt, to infinity. The two functions just presented are closely related by the following relationship:

$$h(t) = -\frac{d}{dt} \log[S(t)] \tag{4}$$

that is, the risk function $h(t)$ can be obtained by deriving the logarithm of the survival function $S(t)$ with respect to time.

Finally, the **cumulative risk function** can be obtained by integrating the risk function

up to time t:

$$H(t) = \int_0^t h(u)du \tag{5}$$

Intuitively, it is defined as the sum of the risk functions for all times before or equal to t. In other words, it represents the cumulative probability of occurrence of the event of interest up to time t.

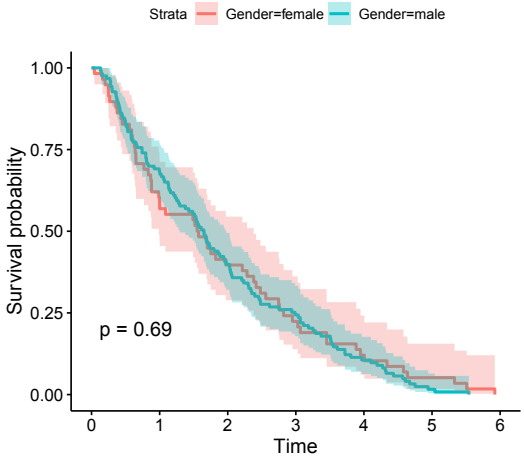

Figure 1: Plot of estimated survival curves for a two-levels factor with 95% confidence intervals and p-value of the Log-Rank test.

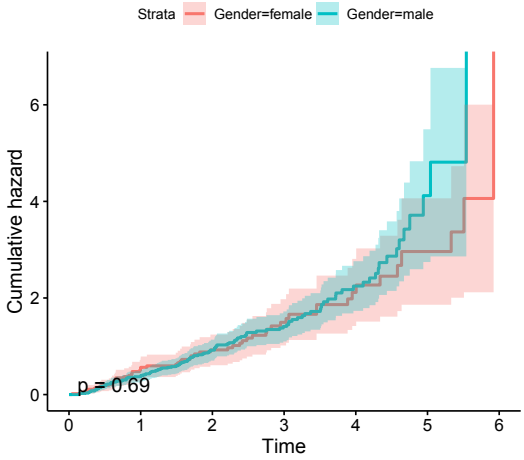

Figure 2: Cumulative hazard plot for a two-levels factor with 95% confidence intervals and p-value of the Log-Rank test.

According to the formulas previously introduced, it is natural to ascertain that the following relations hold: $S(x) = e^{-H(x)}$ and $H(x) = -\log(S(x))$.
Assuming a right censorship mechanism, the survival function can be inferred using the **Kaplan-Meier estimator**:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{6}$$

The quantity $\hat{h_i} = d_i/n_i$ represents the MLE estimate of the risk function, relative to the conditional probability that a subject who is still at risk just before it suffers the event at instant t_i.
The Kaplan-Meier estimator is a non-increasing function, continuous from the right and at intervals, for which the amplitude of the jumps is proportional to $\hat{h_i} = d_i/n_i$, and therefore increases when the number of observed events at t_i increases.
If the last survival time is an event, then $\hat{S}(t) = 0$ from this time onward, while if the last observed survival time is a censoring, then $\hat{S}(t) > 0$.
For a sufficient sample size, the approximation to a normal random variable can be applied, according to which:

$$\hat{S}(t) \dot{\sim} \mathcal{N}(S(t), \hat{V}(\hat{S}(t)) \tag{7}$$

where the variance is estimated via Greenwood's formula[16].

An alternative estimate for the survival function is provided by the **Nelson-Aalen estimator**, which is defined as it follows:

$$\tilde{S}(t) = e^{-\tilde{H}(t)} \tag{8}$$

The estimated risk function $\hat{h}(t)$ corresponds to the amplitudes of the jumps of the estimated Nelson-Aalen curve.
Further insights into the nature and behaviour of these estimators are exhaustively covered by Klein & Klainbaum[17].

1.2 Cox semi-parametric proportional hazards regression model

1.2.1 Assumptions, formulation and profile likelihood

Specification of a model in the analysis of duration data and especially in survival analysis must address the need to define how survival is related to the type of treatment under study or other characteristics; in any case, it is always a matter of specifying how to regress the risk function based on a certain set of covariates. In this section we shall discuss the main characteristics of the Cox model, focusing on the assumptions underlying its structure and formualtion.

On a general level, the Cox model specifies the hazard for an individual i as:

$$h_i(t|Z) = h_0(t)e^{Z_i(t)\beta} \tag{9}$$

where $h_0(t)$ is an unspecified, nonparametric nonnegative function of time called the *baseline hazard*, β is a p-dimensional vector of coefficients and $Z_{ij}(t)$ is the jth regressor of the ith subject (from which the appellation of *semi-parametric*); from this, it's natural to think that Z_i denotes the whole regressor vector for the individual i.
For this model, it's assumed a right-censoring mechanism, also the censoring must be independent from Z.
Setting the covariates as *fixed* with respect to time is imperative to satisfy the core assumption which states that the hazards are, in fact, proportional, whence the explicit form of the Cox model:

$$h_i(t|Z) = h_0(t)e^{Z_i\beta} \tag{10}$$

Insights about the methodologies for verifying this assumption are discussed in Section 1.2.2.
The exponential operator ensures that the final estimates of the outcome of interest are a physical possibility, by implying that the observed deaths (or events) can not

unhappen, meaning event rates can not be negative; the main consequence of this is that covariates have multiplicative effect on baseline risk and, consequently, additive effect on the baseline log-risk[31].

The term *proportional hazards* comes from the assumption that the hazard ratio for two subjects with time-fixed regressor vectors Z_i and Z_j:

$$\frac{h_i(t|Z)}{h_j(t|Z)} = \frac{h_0(t)e^{Z_i\beta}}{h_0(t)e^{Z_j\beta}} = \frac{e^{Z_i\beta}}{e^{Z_j\beta}} \tag{11}$$

is constant over time. Estimation of β is discussed later in this section.

Under these premises, the corresponding survival function can be written in the form of:

$$S(t|Z) = S_0(t)^{e^{\beta^T Z}} \tag{12}$$

where the baseline survival for $Z = 0$ is:

$$S_0(t) = e^{-H_0(t)} = e^{-\int_0^t h_0(u)\,du} \tag{13}$$

For the i-th covariate assumed continuous, e^{β_i} expresses by how much the risk of the event of interest varies multiplicatively for each unit change in Z_i, all other covariates being equal. Meanwhile, if Z is assumed to be categorical, saying there are Z_1, Z_2... Z_l levels, the relative risk e^{β_i} for the variable Z_i expresses how many times the risk that the event is likely to occur increases/decreases. For example, if Z has only two levels, let's say $[0, 1]$, then the relative risk for Z is given by:

$$\frac{h(t|Z=1)}{h(t|Z=0)} = e^{\beta_1} \tag{14}$$

Another important assumption is that there must be no *ties* among the data, that is, no more than one event can happen at time t: ties are usually more likely to occur when the event time scale is discrete or because continuous event times are grouped into intervals. The two possibilities imply different probability structures which are reflected in the approximation that is chosen among different approaches; further insights on this issue are covered by Therneau and Grambsch[31].

The **profile likelihood** is a useful tool for constructing confidence intervals when the maximum likelihood estimates (*MLEs*) of the parameters are of interest.[16]

For the Cox proportional hazards model, the likelihood function can be written as it follows:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right] \tag{15}$$

where β represents the vector of regression coefficients, x_i is the vector of covariates for the ith individual, δ_i is the censoring indicator and $R(t_i)$ denotes the risk set at time

t_i. To obtain the profile likelihood function, we fix the parameter of interest, say β_k, at a particular value, denoted as β_k^0, and maximize the likelihood function with respect to the remaining parameters β_{-k}. This yields the profile likelihood function for β_k:

$$PL(\beta_k) = \max_{\beta_{-k}} L(\beta_k, \beta_{-k}; \beta_k^0), \tag{16}$$

where β_{-k} represents all parameters except for β_k.

1.2.2 Proportional hazards assumption assessment and nonproportionality handling strategies

We focus here on testing the adequacy of the Cox model in relation to the assumption of proportional risks.

If we consider a set of time-fixed covariates with a relatively small number of levels, a useful graphical test for this assumption is to directly take a look at the survival curves: if the assumption holds the log curves should not consistely drift apart[10]. The **Kaplan-Meier curves**, under the definition of survival function presented at (8), exhibit approximately parallel behaviour if plotted on log-log scale. If the variable is continuous or can't be divided in a smaller number of classes because of its many levels, Kaplan-Meier plots do not represent the best option.

Another common method to test the proportional hazards assumption is provided by considering **time-dependent coefficients**, which result in the specification of the model as:

$$h_i(t|Z) = h_0(t)e^{Z_i\beta(t)} \tag{17}$$

If $\beta(t)$ is not constant, the effect of a covariate may consequently not be constant over time. This is the case, for example, when a subject develops resistance in response to a certain treatment, such as an antibiotic. The proportional hazards assumption implies that $\beta(t) = \beta$ and the main consequence of this is that $\beta_j(t)$, if plotted against time, would approximately be described by a horizontal line[11].

Another method to test for proportional hazards is given by considering scaled **Schoenfeld residuals**, which are defined for the kth event as:

$$s_k = \int_{t_{k-1}}^{t_k} \sum_i (X_i - \bar{x}(\hat{\beta}, s)) d\hat{N}_i(s) \tag{18}$$

where $N_i(t)$ comes from the counting martingale process for the ith individual and the d_i is the deviance residual under the same circumstances[6]. The set of Schoenfeld residuals is a p column matrix with one row per event. Grambsch and Thurneau[10] demonstrate that, given $\hat{\beta}$, the coefficient from an ordinary Cox model, then:

$$E(s_{kj}^*) + \hat{\beta}_j \approx \beta_j(t_k) \tag{19}$$

where s_k^* is the scaled **Schoenfeld residual**. From this relation comes the possibility to plot $s_k^* j + \hat{\beta}_j$ against time or a certain function of time $g()$, as a tool to visualize the level of nonproportionality. A line is fitted to the plot followed by a test for zero slope: a nonzero slope is evidence against the hypothesis of proportional hazards. Further insights on the test statistic used under the null hypothesis are discussed by Therneau and Grambsch[31].

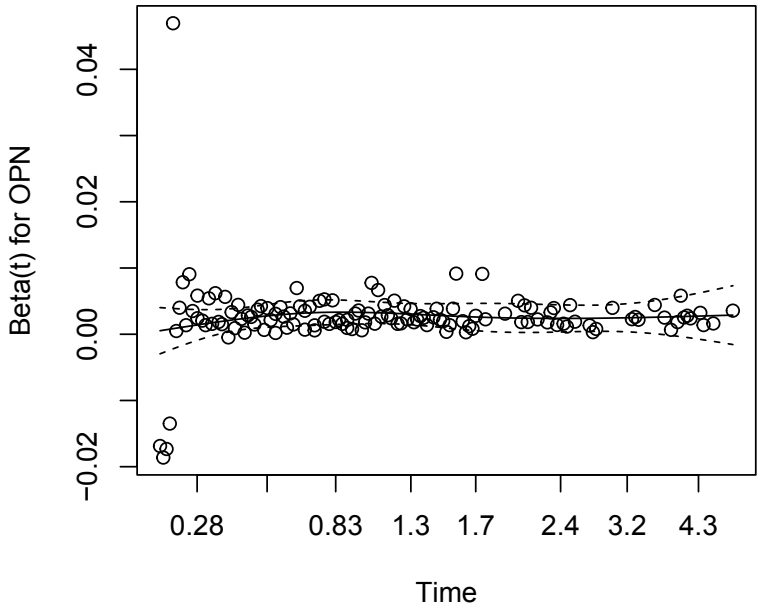

Figure 3: Plot of Schoenfeld residuals v. time.

As mentioned earlier, if we suspect that the effect on the hazard is not constant with time, we can incorporate a time-dependent coefficient, thus making the model result in the following expression:

$$\lambda(t) = \lambda_0(t) e^{Z\beta(t)} \tag{20}$$

If $\beta(t)$ is not constant, the impact of one or more regressors may vary over time. To verify this, it's possible to plot $\beta_j(t)$ against time: if the impact on the hazard is not time-dependent, this will result in a horizontal line.
If nonproportionality of the hazards does in fact appear during the analysis of the data, there are a certain number of strategies, depending from case to case, one can apply to possibly overcome the issue:

- **Stratification**: It's possible to incorporate covariates which cause violation of the proportional hazards assumption as stratification factors rather than considering them as regressors. The shape of the baseline risk varies for each category while

the effect of covariates remains equal in all strata. This means that the stratified model is *not* equivalent to estimating l separate models for each category. Inference is performed by constructing l functions of partial likelihood, each specific to the respective layer.
While this approach is of relatively easy to use, it comes at the cost of some drawbacks, which are:

  - Lose of information of the variable used as a single factor to perform the stratification on the overall survival since it is omitted from the set of regressors.
  - While stratification comes as a natural possibility for categorical variables, quantitative variables have to be discretized into intervals, but choosing how many and how wide said intervals are is not obvious and can result in bias for the coefficients of the regressors or a diminishment in efficiency.
  - Stratified analyses are, in general, less efficient compared to the ones without stratification factors or to anaylses which include interactions with time, when a time-dependent structure is assumed to be present.

- **Use of time-dependent covariates**: Time-dependent covariates are a possibility when working with time-varying effects in a sense that a time-dependent covariate $X(t)$ can be created so that:

$$\beta(t)X = \beta X(t) \tag{21}$$

  The choice for $X(t)$ has to be considered in relation to the specific goal of the study and to the relative theoretical considerations or it can be a function chosen as evidence emerged from the smoothed residual plots. In section 1.2.3 we provide an overview on the usage of P-Splines which serve as a very flexible tool to model the functional form of a covariate.
  Frequently, time-dependent covariates are a repeated measure of a certain variable over the period of observation. In these cases, there might be correlation among the single measures, for example when multiple doses of a drug are administered over time. A common approach is to define (start, stop] intervals for the variable, thus assuming a step function which jumps at the measurement points. The flexibility of the (start, stop] approach is frequently used in survival analysis and, more in general, in EHA (event history analysis), however sometimes choosing the points of break of the intervals might not be so obvious, especially when there are more than two measurements; short intervals might result in producing biased estimates as well.

- **Use of AFT or (multiplicative-)additive hazards models**: There are cases in which the data are more suited to be modeled with Accelerated Failure Time (AFT) models, which are common in industrial applications. In survival analysis'

context, the primary goal is to study the time until an event of interest occurs, such as death, failure, or recurrence of a disease.
The AFT model assumes that the logarithm of the survival time, denoted as T, follows a linear relationship with covariates. The general formulation of the AFT model is given by:

$$\ln(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \sigma\epsilon \tag{22}$$

where:

- $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients for the covariates $x_1, x_2, \ldots, x_p$, respectively.
- σ is a positive parameter representing the scale parameter of the model.
- ϵ is a random error term assumed to follow a standard extreme value distribution with a location parameter of 0 and a scale parameter of 1.

The AFT model implies that a one-unit increase in any of the covariates multiplies the survival time by a constant factor. If β_j is the regression coefficient of a covariate x_j, then the acceleration factor is given by e^{β_j}. This factor determines the amount of acceleration or deceleration of the survival time for each unit change in the corresponding covariate. For instance, if the acceleration factor is 2, it means that a unit increase in the covariate results in a doubling of the survival time, indicating a decelerating effect. Conversely, an acceleration factor less than 1 would lead to a shorter survival time for higher values of the covariate, implying an accelerating effect. More insights on AFT models are provided by Klein[16].
Another approach, even more flexible in at least some cases, is given by working with two whole different families of models which consider the effects of the covariates on the risk respectively on a additive and a multiplicative-additive scale; these two categories of models for the hazard function are presented and discussed as core of this work, respectively, in sections 2.2 and 2.3.

- **Checking for omitted variables**: While we'll just provide a mention of this other issue, it is also worth noting that omitting covariates can, in some cases, be cause for a lack of proportionality among the hazards: we illustrate this scenario by considering a simple model in the form of:

$$\lambda(t) = \lambda_0 e^{x_1\beta_1 + x_2\beta_2} \tag{23}$$

given that x_1 is a 0-1 binary treatment indicator and x_2 is an important predictor. By omitting x_2 for any reason from the initial model we would fit it as:

$$\lambda(t) = \lambda_0 e^{x_1\beta} \tag{24}$$

This could result in the violation of the assumption of proportional hazards and also in the fact that the partial likelihood estimate of β based in the misspecified model is a biased estimate of β_1 since, when a covariate is ignored, the operative hazard is the average hazard of those at risk at each time point, a mixture of hazards[31].

1.2.3 Evaluation methods for the functional form and P-Splines usage

As previously stated, when the proportional hazards assumption is satisfied, this implies that, considering a variable related to the age of the subjects as an example, the ratio of the risks between a 30- and a 45-year-old is the same as that between a 65- and a 80-year-old. What if the risk does not begin to rise until a certain age or if it exhibits a non-constant behaviour throughout the years? Meaning, we are trying to understand if some sort of *nonlinearity* is present in the effects of a variable on the risk over time. This issue is consistent with the necessity of investigating correct procedures for evaluating the best functional form for a covariate; we'll show two main procedures: martingale residuals plots and P-splines.

As for the first method, we firstly have to define what **martingale residuals** are: let T_i be the survival time of the i-th subject, and let $\hat{H}(t)$ be the estimated cumulative hazard function at time t. The martingale residual r_i for the i-th subject is computed as:

$$r_i = \frac{\delta_i - \hat{H}(T_i)}{\sqrt{\hat{V}(\hat{H}(T_i))}} \tag{25}$$

where δ_i is the event indicator variable for subject i (1 if an event occurred, 0 otherwise), and $\hat{V}(\hat{H}(T_i))$ is the estimated variance of the cumulative hazard function at time T_i.

Nonlinearity is not an issue for categorical variables, so we only examine plots of martingale residuals against a continuous variable.

Therneau et al.[31] suggest plotting the martingale residuals from a null Cox model, where $\hat{\beta} = 0$, against each variable separately and superimposing a scatterplot smooth (an example is shown in Figure 4). Note that fitted lines should be approximately linear to satisfy the Cox proportional hazards model assumptions.

The interpretation of this type of graph is very close to that of scatterplots used to assess the relationship between one response variable and another covariate in the usual linear regression models for data that do not have censoring.

In the software R, the procedure can effectively be implemented by using the function **`ggcoxfunctional()`** of the library **`survminer`**, more details related to this command can be found at the R *help* section[19].

The other approach to identify the most appropriate functional form of a covariate is via **P-splines**.

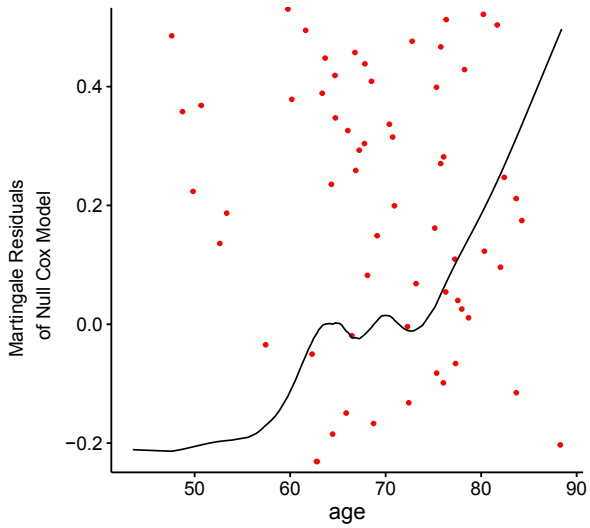

Figure 4: Martingale residuals from a null Cox model plotted against a continuous variable inherent to age.

On a general level, splines are a mathematical technique used to approximate complex functions or irregular relations between two or more variables. A spline is a piecewise curve composed of low-degree polynomial segments, connected in a way that ensures continuity and smoothness of the resulting curve[13]. The main goal of splines is to provide an accurate approximation of a dataset without being excessively influenced by noise or random fluctuations.
A spline of degree k with n nodes can be defined as:

$$S(x) = \sum_{i=1}^{n} c_i N_i(x) \tag{26}$$

where:

- $N_i(x)$ are the *basis functions* (often polynomials) that depend on the nodes and form the piecewise curve.
- c_i are the coefficients that determine the height of each segment.

The choice of basis functions $N_i(x)$ is often made to ensure the continuity and smoothness of the curve. A common example of a spline is the cubic spline, where each segment is a cubic polynomial.
P-splines (penalized splines) are a variant of splines that use a regression approach to estimate the coefficients of the basis polynomials. Unlike traditional splines, p-splines incorporate a penalty term in the optimization process to control the complexity of the model and reduce overfitting[23].
P-splines are particularly useful when dealing with large amounts of data or when tighter control over model flexibility is desired. Additionally, p-splines can be compu-

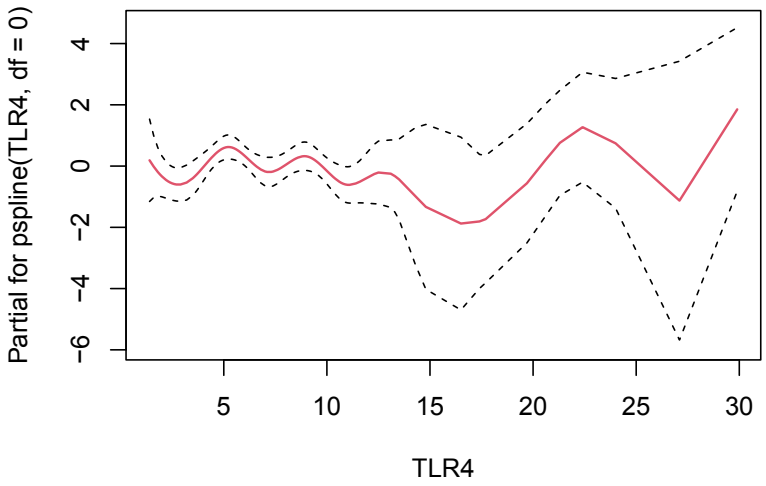

Figure 5: P-spline interpolating a variable.

tationally more efficient than traditional splines.
In the case of p-splines, the model is defined as:

$$S(x) = \sum_{i=1}^{n} c_i N_i(x) + \lambda \int S''(x)^2 \, dx \tag{27}$$

where:

- The first term is similar to that of traditional splines, with coefficients c_i estimated through regression analysis.
- The second term is the penalty term, penalizing the curvature of the spline. λ is a regularization parameter controlling the intensity of the penalty.

The most appropriate number of degrees of freedom used to approximate the form of a variable can be computed by several optimization methods such as cross-validation or automatic methods based on Akaike's Information Criteria; the correct number of degrees of freedom represents a compromise between variance and bias, as if it is too low the spline won't be able to explain the behaviour of data, and if too high, as previously stated, it will result in overfitting.
While the usage of P-splines provides a simple yet powerful tool of the functional form analysis of a covariate, this procedure suffers from the additional technical difficulty of nonpredicatability: the value of the smooth at any point in time i is a function of resiuduals from the future relative to that point[31].
In the software R, smoothing splines can be used via the package **`survival`** by using the command **`pspline()`**[28]. The command can be used on the regressors of a

Cox model; note that setting the argument **df** to 0 implies choosing AIC as an automatic method to select the number of degrees of freedom for the variable on which **pspline()** is used.

2 Extending the Cox model and flexible regression

In this section, after introducing the **extended Cox model for time-varying effects** and the relative procedure to test for their significance, we explore the family of **additive hazards models**; said models relate the conditional hazard function of the failure time to covariates in a linear way. The relation between risk and regressors is expressed in form of a risk difference rather than a risk ratio, which was the case for the multiplicative hazards model.
The additive hazards framework can be used effectively to incorporate frailty and to handle interval-censored data, and the semi-parametric inference deriving from its structure results in much simpler inference procedures computationally speaking.[20]
Last, we focus on the class of **multiplicative-additive hazards models**, exploring in particular the Cox-Aalen model, which, as we'll show, represents a compromise between the first two classes of models presented in this dissertation.

2.1 Extended Cox model for time-varying effects

As mentioned in the previous sections, there are ways in which, to a certain extent, variables with time-varying effects can still be incorporated in the Cox model by relaxing its assumptions. To do so, we now consider the more general **Cox model with time-varying regression effects**.
Let's allow the coefficients of the model to be able to depend on time, thus resulting in the following formulation:

$$h(t|Z) = h_0(t)e^{Z(t)\beta(t)} \tag{28}$$

$\beta(t)$ is now a vector of regression functions which depend on time. Ideally, to best explain the behaviour of the phenomena over time, it's often better to consider the parametric part of the model as split in two parts, one whose effects are time-dependent and one whose are not, thus resulting in the following expression:

$$h(t|Z) = h_0(t)e^{Z_1(t)\beta(t)+Z_2(t)\gamma} \tag{29}$$

Parameters can still be estimated via partial likelihood.
To investigate if the coefficients included in the model as time-dependent are, in fact, dependent on time, we need to test if their effects on the risk are constant; to do

this, it is necessary to provide an estimation of γ and $B(t) = \int_0^t \beta(u)du$; the quantity $\sqrt{n}(\hat{B}(t) - B(t))$ is asymptotically a gaussian process with mean equal to zero, this can be used to construct uniform confidence bands.[22] The hypothesis $H_0 : \beta_1(t) = \eta_1$ is validated by obtaining a p-value derived from a supremum test statistics defined as it follows:

$$\sup_{t\in[0,\tau]} |\hat{B}_1(t) - \frac{\hat{B}_1(\tau)}{\tau} t| \tag{30}$$

One can consider, as a simple approach, to establish, at the beginning, all coefficients of covariates as time-dependent, thus excluding from the model those which doesn't significantly appear as nonlinear, and then re-fit the model with only covariates which resulted as significantly nonlinear upon performing said test. Since it's not always easy to establish the correct functional form for these covariates, it can be useful to consider automatic procedures to evaluate the shape of the estimates with techniques as P-splines, as shown in section 1.2.3.

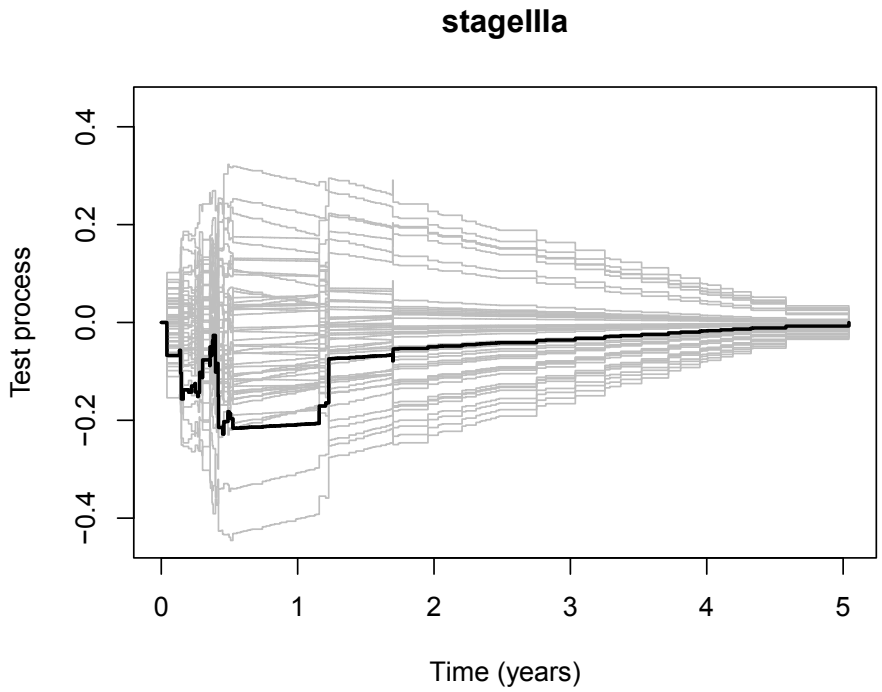

Figure 6: Observed test-process for a covariate along with 50 simulated processes under the null hypothesis of proportional hazards.

The estimation of the survival function can be somehow complicated since we have a fixed set of regressors in the model; on a general level, given a set of covariates Z_0, said function can be expressed in the form of:

$$e^{\int_0^t \lambda_0(u) e^{Z_0 1 \beta(u)} du} \tag{31}$$

which requires estimates of $\beta(t)$ and γ. The difficulty of working with such quantity is given by its computational onerousness, further details on the exact procedure are provided by Martinussen & Scheike[22]. A graphical representation of the effects over time on the risk for each covariate can be obtained by simulating a certain number of processes under the null hypothesis of time-invariant effects; said approach is quite useful for determining at which point in time the process deviates from the time-indipendency assumption's zero line.[22]
Focusing on the inferential side, while in the proportional hazards model all the time-varying regression coefficient are constant (under the hypothesis $H_0 : \beta(t) = \beta$), now the regression coefficients are considered individually and one can investigate the two hypotheses which follows:

$$\begin{cases} H_{01} : \beta_p(t) = 0 \\ H_{02} : \beta_p(t) = \beta_p \end{cases} \tag{32}$$

so we can focus on the p-th regression coefficient without loss of generality. It is important to notice that the other regression coefficients in the model are allowed to vary with time.
Testing the significance of the regression coefficients will equivalently lead to construction of confidence bands; we address the reader to Martinussen & Scheike[22] for a deeper overview on these testing procedures.

2.2 Additive hazards models

2.2.1 Aalen's additive hazards model

Aalen's additive hazards model is a fully nonparametric model expressed in the following form:

$$h(t|Z) = Z^T \alpha(t) \tag{33}$$

where $\alpha(t)$ is the regression coefficients vector composed by functions representing time-varying effects of the covariates on the base risk over the time. The first term of Z will usually be 1, while the term $\alpha_1(t)$ represents the baseline hazard; every other term included in the model, $Z_j \alpha_j(t)$, refers to the excess additive risk due to the presence of the j-th regressor with respect to the baseline risk.
The inferential procedures can be easily performed by evaluating the **cumulative regression coefficients** which are defined as:

$$A(t) = \int_0^t \alpha(t) du \tag{34}$$

whose estimation is given via least squares. More details on the estimation procedure are provided by Klonecki et al.[18].

There can be a certain variety of hypotheses about the regression coefficients which can be evaluated; we therefore show a test-statistic based on the estimated cumulative regression coefficients to investigate said hypotheses.
Cumulative coefficients are better suited than regression coefficients when it comes to inference for the additive hazards model. $\beta(t)$ will have a bias part and variance part.[22]
In the following, we consider the two hypotheses:

$$\begin{cases} H_{01} : \beta_p(t) = 0 \\ H_{02} : \beta_p(t) = \gamma \end{cases} \tag{35}$$

meaning that, without loss of generality, we formulate the hypothesis for the p-th regression coefficient function. Both these hypotheses are about the functional behavior of the regression coefficient function and the stated equalities are for the entire considered time range $[0, \tau]$. These hypotheses may also be of relevance for multiple regression coefficients simultaneously and all the procedures mentioned here can easily be generalized to a multivariate setting.
We now switch from the hypotheses above to the respective ones involving the cumulative regression coefficients defined at (34):

$$\begin{cases} H_{01} : B_p(t) = 0 \\ H_{02} : B_p(t) = \gamma t \end{cases} \tag{36}$$

Again, it is possible to consider a maximal deviation test statistic such as:

$$\sup_{t \in [0,\tau]} |\hat{B}_p(t)| \tag{37}$$

If $B_p(t)$ is expected to be monotone, it's possible to use $B_p(\tau)$ to test the null hypothesis, but this test statistic will have low power if $B_p(\tau)$ is equal to zero. On the other hand the test statistic will have low power if $\beta_p(t)$ differs only substantially from 0 towards the end of the time period $[0, \tau]$; given so, a test statistic in the form of:

$$\sup_{s,t \in [0,\tau]} |\hat{B}_p(s) - \hat{B}_p(t)| \tag{38}$$

should be better in terms of detecting departures of $\beta_p(t)$ from the null hypothesis. Further insights on this matter are exhaustively covered by Martinussen & Scheike[22].
The relative survival probability for the model is given as:

$$e^{X_0^T B(t) - Z_0^T B t \gamma} \tag{39}$$

2.2.2 Ling & Yin's and McKeague & Sasieni's semi-parametric additive hazards models

As stated in the previous section, all coefficients in the Aalen model are time-varying; once it's been fitted, and after having observed which covariates do in fact exhibit time-varying effects and which don't, it may be of interest to split the model in two parts, one with time-varying coefficients and one with constant effects. The model resulting from this diagnostic process of time-varying coefficients selection is the **semiparametric additive hazards model of McKeague & Sasieni**:

$$h(t) = Z^T\alpha(t) + X^T(\gamma) \tag{40}$$

in which γ, the effects of the set of regressors X, are fixed in time. Given this, it comes natural to think of the Aalen model as a special case of the McKeague & Sasieni's.
The model represents a good compromise in terms of flexibiliy, variance and bias when there's indeed a heterogeneity among the behaviours of the effects of the covariates over time.
When p, the dimension of X, is equal to 1, the model can be written as follows:

$$h(t) = Z^T\alpha(t) + \beta(t) \tag{41}$$

and is known in survival analysis literature as the **Lin & Ying's semiparametric additive hazards model**[20].
From a practical point of view it is preferable to work with a more elaborate model that can describe time-dynamics of covariate effects when needed, rather than forcing all regression effects to be constant.
In section 2.2.1 we provided a simple test to verify if a covariate effect was significant and to test if a covariate had a time-invariant effect using the full Aalen additive model. The test was limited to considering one covariate only, and although it's possbile to construct a multidimensional version of it, it is often preferable to work with successive tests for time-varying effects, that is testing one component at a time using the reduced model as the starting point for the next analysis and test.
As for inference, we shall focus on the two hypotheses:

$$\begin{cases} H_{01} : B_p(t) = 0 \iff B_p(t) = 0 \\ H_{02} : B_p(t) = 0\gamma_{q+1} \iff B_p(t) = \gamma_{q+1}t \end{cases} \tag{42}$$

where, without loss of generality, we consider only the last nonparametric component of the model.
Focusing on H_{01}, a confidence band for $B_p(t)$ can be obtained (analytical details are provided by Martinussen & Scheike[22]). In the following, we illustrate a resampling

approach to provide a well described limit-distribution for $n^{1/2}(\hat{B}_p(t)-B_p(t))$. The resampling approach is based on the following decomposition into i.i.d. residuals. First, note that:

$$n^{1/2}(\hat{\gamma}-\gamma) = C_1^{-1} n^{-1/2} \sum_{i=1}^{n} \epsilon_{2i} + o_p(1) \tag{43}$$

where:

$$C_1 = n^{-1} \int_0^{\tau} Z^T(t) H(t) Z(t) dt \tag{44}$$

given:

$$H(t) = W(t)(I - X(t)X^{-}(t)) \tag{45}$$

in which $W(t)$ is a diagonal weight matrix. Martingale residuals are then derived as:

$$M_i(t) = N_i(t) - \int_0^{\tau} Y_i(s)(X_i^T(s)dB(s) + Z_i(s)\gamma ds) \tag{46}$$

The sum of martingales is asymptotically equivalent to a sum of i.i.d. terms:

$$\tilde{\epsilon}_{2i} = \int_0^{\tau} (Z_i(t) - \mathbf{E}(Y_i(t)Z_i(t)X_i^T(t))\mathbf{E}(Y_i(t)Z_i(t)X_i^T(t))^{-1}X_i(t))dM_i(t) \tag{47}$$

An estimation of the variance of $n^{1/2}(\hat{\gamma}-\gamma)$ is provided by:

$$C_1^{-1}(n^{-1}\sum_{i=1}^{n} \hat{\epsilon}_{2i}^{\otimes 2})C_1^{-1} \tag{48}$$

and where $\hat{\epsilon}_{2i}$ is estimated using:

$$\hat{M}_i(t) = N_i(t) - \int_0^{\tau} Y_i(s)(X_i^T(s)d\hat{B}(s) + Z_i^T(s)\hat{\gamma} ds) \tag{49}$$

the variance of $n^{1/2}(\hat{B}_p(t)-B_p(t))$ can be estimated with the robust variance estimator:

$$\hat{\psi}(t) = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_{3i}^{\otimes 2}(t) \tag{50}$$

of which:

$$\epsilon_{3i}(t) = \epsilon_{4i}(t) - P(t)C_1^{-1}\epsilon_{2i} \tag{51}$$

noting that:

$$\epsilon_{4i}(t) = \int_0^{\tau} (n^{-1}X^T(s)X(s))^{-1}X_i(s)dM_i(s) \tag{52}$$

and:

$$P(t) = \int_0^{t} X^{-}(s)Z(s)ds \tag{53}$$

The relative estimated survival function for the semi-parametric model is given by:

$$\hat{S}_0(t) = e^{-X_0^T \hat{B}(t) - Z_0^T t\gamma} \tag{54}$$

Further details on its pointwise confidence intervals are deeply discussed by Martinussen & Scheike.[22]
One drawback in using additive hazards models is that estimates of the regression coefficients may be negative, as a consequence of this their cumulative estimates will therefore decrease. One alternative, as we show in the following section, is to consider a further generalisation to overcome the issue, which is, in this context, provided by the family of multiplicative-additive hazards models.

2.3 Multiplicative-additive hazards models

Multiplicative-additive hazards models are born with the aim of incorporating the flexibility of additive models, which are fit to handle time-varying effects, and proportional hazards model based on a multiplicative scale of the effects on the base risk. In this section, we show the main representative of this class of models: the **Cox-Aalen model**.

2.3.1 Cox-Aalen's multiplicative-additive hazards model

The Cox-Aalen model is defined as it follows:

$$h(t) = (X^T(t)\alpha(t))e^{Z^T\beta} \tag{55}$$

For this model, some covariate effects are believed to result in multiplicative effects, whereas other effects are better described as additive. On a practical level, covariates which do exhibit time-varying behaviour would be inserted in the additive part, while the time-constant ones would be incorporated in the multiplicative part of the model.
The main quantities which characterize its formulation, β and $A(t) = \int_0^t \alpha(u)du$ are still easily computed and estimated in order to establish their asymptotic properties.[22]
Since the survival function directly depends on $A(t)$, it is still quite easy task to estimate the quantity:

$$e^{-Z_{01}^T A(t) e^{Z_{02}^T \beta}} \tag{56}$$

Goodness-of-fit procedures, to establish whether the model is appropriate to fit the data or not, are provided upon application of martingale residuals.[22]
Focusing on the inference, the log-likelihood function of the model is derived as:

$$l(\beta) = \sum_{i=0}^{n} \int_0^\tau \log(Y_i(t)X_i(t)^T dA(t) e^{Z_i(t)^T\beta}) dN_i(t) - \sum_{i=0}^{n} \int_0^\tau Y_i(t) e^{Z_i(t)\beta} X_i(t)^T dA(t) \tag{57}$$

where $Y_i(t)$ is the at risk indicator. Therefore, we obtain the score equations inherent to β and $dA(t)$:

$$\begin{aligned} \int Z(t)^T(dN - Y(\beta,t)dA(t)) = 0 \\ Y(\beta,t)^T W(t)(dN - Y(\beta,t)dA(t)) = 0 \end{aligned} \tag{58}$$

where W is a diagonal weighted matrix. If β is known, the estimator of the cumulative intensity is:

$$\hat{A}(\beta,t) = \int_0^t Y^-(\beta,s)dN(s) \tag{59}$$

where:

$$Y^-(\beta,t) = (Y(\beta,t)^T W(t) Y(\beta,t))^{-1} Y(\beta,t)^T W(t) \tag{60}$$

is a weighted generalized inverse of $Y(\beta,t)$ with the convention that $Y^-(\beta,t)$ is 0 when the above inverse does not exist. Inserting this estimator into the score equation for β gives $U(\beta) = 0$ with:

$$U(\beta) = \int_0^t Z^T(t) G(\beta,t) dN(t) \tag{61}$$

in which:

$$G(\beta,t) = I - Y(\beta,t)Y^-(\beta,t) \tag{62}$$

is the projection onto the orthogonal space spanned by the columns of $Y(\beta,t)$. Now, to compute the estimator, we need weights that do not depend on the unknown baseline intensities. Such can be derived as:

$$w_i(t) = Y_i(t) h_i(t) e^{-Z_i(t)^T \beta} \tag{63}$$

in which $h_i(t)$ for $i = 1, ..., n$ are known functions independent from β. $\hat{\beta}$ is given as solution to the score equation:

$$U(\beta,t) = 0 \tag{64}$$

and estimate $A(t)$ by:

$$\hat{A}(\hat{\beta},t) \tag{65}$$

These estimates can now be used to estimate the likelihood weights, the estimators related to such weight are efficient. For further details on the inferential procedures concerning the Cox-Aalen model we refer the reader to Martinussen & Scheike.[22]

3 Application to NSCLC data

Lung cancer is among the most deadly cancers for both men and women. Its death rate exceeds that of the three most common cancers (colon, breast, and pancreatic) combined. Over half of patients diagnosed with lung cancer die within one year of diagnosis and the 5-year survival is around 17.8%.[15]
NSCLC stands for **Non-Small Cell Lung Cancer**, which is the most common type of lung cancer. NSCLC is further categorized into different typologies or subtypes based on the specific cell types and histological features present in the tumor.
Subtypes of NSCLC include:

- **Adenocarcinoma**: Adenocarcinoma is the most common subtype of NSCLC, accounting for about 40% of all lung cancers. It arises from the cells that line the smaller airways and tends to develop in the outer regions of the lungs. Adenocarcinoma is more common in non-smokers and is often associated with certain genetic mutations.
- **Squamous Cell Carcinoma**: Squamous cell carcinoma, also known as epidermoid carcinoma, accounts for approximately 25-30% of NSCLC cases. It typically arises in the central part of the lungs and is strongly linked to smoking. Squamous cell carcinoma develops from the cells that line the bronchial tubes.
- **Large Cell Carcinoma**: Large cell carcinoma is a less common subtype of NSCLC, comprising about 10-15% of cases. It is called *large cell* because the cancer cells appear large and undifferentiated when viewed under a microscope. Large cell carcinoma can occur in any part of the lung and tends to grow and spread rapidly.

The therapeutic approach includes radiotherapy, chemotherapy or a mix of both. Radiotherapy uses high-energy beams to damage DNA within cancer cells, thereby destroying them. This therapy can help control or eliminate tumors at specific sites in the body. Patients with NSCLC that is localized to the chest and who are not candidates for surgical resection may benefit from radiotherapy. Radiotherapy also can be part of palliative care to improve quality of life in NSCLC patients who do not respond to surgery or chemotherapy[1]. As for chemotherapy, the American Society of Clinical Oncology

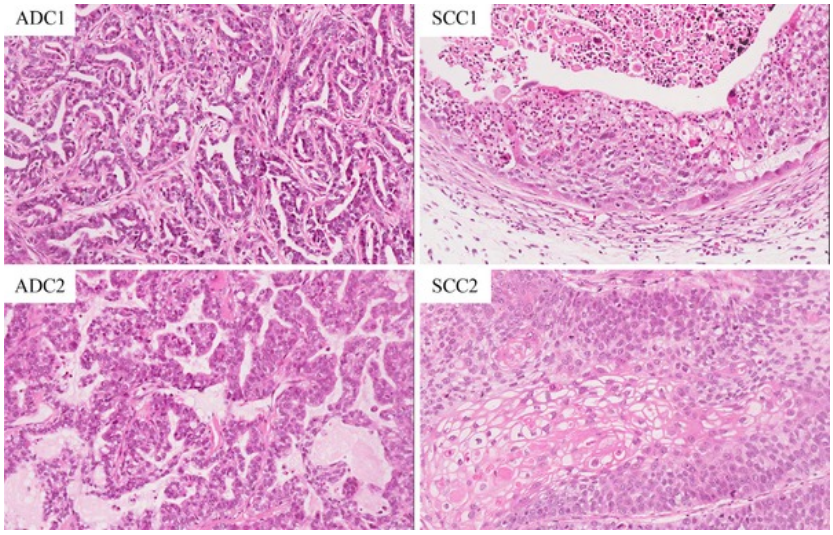

Figure 7: NSCLC histology. HE stain of primary NSCLC PDX tumors including two ADC and two SCC.

states that treatment for a patient with a Performance Status[27] of 0 or 1 is a regimen of a platinum (cisplatin or carboplatin) plus paclitaxel, gemcitabine, docetaxel, vinorelbine, irinotecan, or pemetrexed[24].
Personalized medicine by targeting appropriate molecular targets in tumors has helped improve survival in patients with NSCLC, along with hematic biomarkers which can serve as prognostic factors for the expected survival of the patients, such as in this case. We therefore analysed the dataset described in the following section to show how non-proportional hazards models can serve as an additional tool to model said prognostic factors even when the proportional hazards assumption of the Cox model, starting point of every survival modeling approach, happens to fail, and how the estimated survival can differ noticeably from one model to another, thus leading, consequently, to different conclusions.

3.1 Data overiew and missing values imputation

The following analyses were performed in the software R, flexible models have been fitted using the **timereg**[22] package.
Data are inherent to a cohort of 181 NSCLC patients treated between March 2007 and September 2013. Measurements included biological characteristics such as age and various health-performance indexes, including the following blood-biomarkers: OPN, CA-IX, IL-6, IL-8, CRP, CEA, CYFRA 21-1, VEGF, α2M, TLR4 and sIL2R.
37.5% of patients received radiotherapy alone according to the August 2005 protocol, with an individualized total dose delivered in fractions of 1.8 Gy twice daily, limited by the mean lung dose or the spinal cord dose[32].
55% received concurrent chemo-radiotherapy scheme for a prescribed dose of 45 Gy,

followed by an individualized dose ranging from 8 to 24 Gy, delivered in fractions of 2.0 Gy once daily, again limited by the dose to surrounding organs at risk[33].
6.6% of patients followed the Phase II Positron Emission Tomography (PET) boost trial, in which a dose escalation protocol was based on the Fluorine-18-Fluorodeoxyglucose distribution of the PET scans[34].
We now provide an overview on each of the variables included in the dataset:

- **ID**: Numeric identifier indicating each patient as a statistical unit.
- **Survival**: Quantitative variable which indicates *when*, over the time of the study, each patient has experienced the event of interest, which in this case was the death of the patient itself.
- **Status**: Status indicator which indicates *if* the patient has or has not died over the time of the study.
- **Gender**: Two-levels factor indicating the sex of each patient: *male* or *female*.
- **age**: Discrete quantitative variable indicating the age in years of each patient.
- **stage**: Four-levels factor indicating the state of the disease: levels are *I*, *II*, *IIIa* and *IIIb*.
- **histology**: Three-levels factor indicating the NSCLC histotype: *adeno* is for adenocarcinoma, *SCC* is for squamous cells carcinoma and finally *NOS* stands for cell lung carcinoma not-otherwise specified.
- **WHO-PS**: Stands for *World Health Organization Performance Status* and serves as an indicator which quantifies the capability of a subject to carry out all normal activities without restriction. It is here represented as a five-level factor with levels: 0, 1, 2, 3, 4. The levels stand for:
  - **0**: Able to carry out all normal activity without restriction.
  - **1**: Restricted in physically strenuous activity but ambulatory and able to carry out light work.
  - **2**: Ambulatory and capable of all self care but unable to carry out any work; up and about more than 50% of waking hours.
  - **3**: Capable only of limited self care; confined to bed or chair more than 50% of waking hours.
  - **4**: Completely disabled; cannot carry out any self care; totally confined to bed or chair.
- **FEV1s**: Continuous quantitative variable indicating the forced expiratory volume in one second for a patient.

- **Lymph nodes**: Discrete quantitative variable indicating the number of positive lymph node stations identified in the diagnostic PET scans.
- **RT Protocol**: Type of protocol corresponding to the treatment of each patient. Type of protocol is encoded in a three-level factor as: *Concurrent RT*, *PET Boost*, *New protocol August 2005*, i.e. standard external beam radiation therapy (EBRT).
- **Total dose (1st)**: Continuous quantitative variable indicating the Gy dose received in the radiotherapy.
- **Total dose (2nd)**: Continuous quantitative variable indicating the additional Gy dose received in the second exposition to radiotherapy.
- **GTV**: Continuous quantitative variable indicating gross tumour volume[5].
- **OPN**: Discrete quantitative variable related to *osteopontin*, a blood-biomarker related to hypoxia. More info on this protein are provided by Lund et al[2].
- **CA-9**: Discrete continuous variable related to the blood-biomarker *carbonic anhydrase IX*[14].
- **IL 6**: Continuous quantitative variable related to the level of *interleukin-6*, blood-biomarker related to inflammation response[9].
- **IL 8**: Continuous quantitative variable related to the level of *interleukin-8*, blood-biomarker related to inflammation response[29].
- **CRP**: Continuous quantitative variable related to the level of *C-reactive protein*[7].
- **CEA**: Continuous quantitative variable related to the level of *carcinoembryonic antigen*, a blood-biomarker inherent to tumour load[12].
- **Cyfra 21-1**: Continuous quantitative variable related to the level of *cytokeratin fragment*, a blood-biomarker inherent to tumour load[35].
- **alpha-2M**: Continuous quantitative variable related to the level of *carcinoembryonic antigen*, a blood-biomarker inherent to immunologic response[4].
- **sIL2R**: Discrete quantitative variable related to the level of *carcinoembryonic antigen*, a blood-biomarker inherent to immunologic response[25].
- **TLR-4**: Continuous quantitative variable related to the level of *toll-like receptor 4*, a blood-biomarker inherent to immunologic response[8].
- **VEGF**: Continuous quantitative variable related to the level of *vascular endothelial growth factor*, a blood-biomarker inherent to immunologic response[26].

The initial situation of the dataset, for those variables that had NA values, is shown in Table 1. An imputation of missing values was chosen in order to handle variables which showed a percentage of NAs/observations ratio below 10%, since literature suggests that, below this threshold, estimates obtained via imputation can still be considered reliable[3]. Imputation was performed by replacing the ith missing data with the mean of the values if there was no evidence of skewness in the distribution of that covariate, viceversa with the median if a varying degree of skewness was, indeed, present. Mode imputation was chosen for categorical variables with a percentage of NAs/Observations <10%.

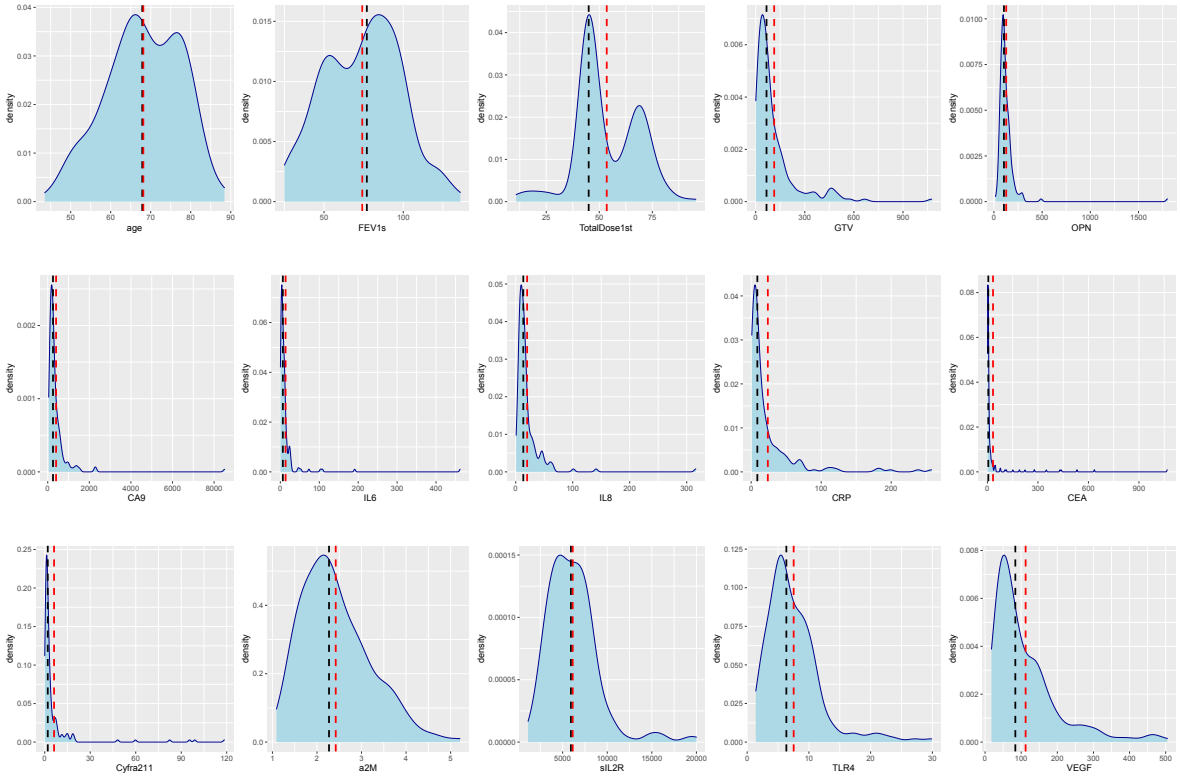

Figure 8: Density graphs of quantitative variables. The median is indicated by the black dashed line, the mean by the red one.

To handle variables with more than 10% NAs/observations, one approach consists in using the predicted values obtained via linear regression for the variable against all other predictors, which could have been the case for FEV1s; however, since the output of the MCAR test[21] for this covariate was significant (t_{oss}: 36.0, df: 23, p: 0.0408), a more robust approach was chosen, i.e. **multiple imputation by chained equations**, commonly known in the statistical literature as MICE[3]. Another issue arisen during the phase of data inspection was that, for some covariates, values under a certain threshold were not reported; this is most likely due to the fact that the biomarkers could not be detected by the biomedical equipment under said limit. In the original

Variable	NAs	Type	Imputation/Motivation
TLR-4	1	Quantitative	Median imputation: Skewed distribution.
VEGF	1	Quantitative	Median imputation: Skewed distribution.
Cyfra 21-1	1	Quantitative	Median imputation: Skewed distribution.
OPN	1	Quantitative	Median imputation: Skewed distribution.
CA-9	1	Quantitative	Median imputation: Skewed distribution.
Stage	2	Qualitative	Mode imputation: Qualitative, NAs/data < 10%.
GTV	3	Quantitative	Median imputation: Skewed distribution.
alpha2M	5	Quantitative	Median imputation: Skewed distribution.
histology	8	Qualitative	Mode imputation: Qualitative, NAs/data < 10%.
IL-8	11	Quantitative	Insertion of arbitrary values between 0 and x for value < x.
WHO-PS	12	Qualitative	Mode imputation: Qualitative, NAs/data < 10%.
IL-8	21*	Quantitative	Insertion of arbitrary values between 0 and x for value < x.
FEV1s	35*	Quantitative	MICE (MCAR test: $t_{oss} : 36.0, df : 23, pval : 0.0408$).
TotalDose(2nd)	78**	Quantitative	Excluded from imputation for excess of NAs.

Table 1: Number of NAs for each variable with one or more NA values along with the type of Missing Values Imputation performed and relative motivation. *: NAs/data >10%. **: NAs/data >30%.

dataset, these values were reported simply as being inferior to a certain value. Given the nature of these hematic components, and the fact that obviously their presence in the circulatory system can not be negative, given also that the threshold was, for each covariate, very low, imputation was made in this case by simulating values from a uniform distribution defined in (0, m], where m is the threshold of detection.
It was decided to remove the covariate TotalDose(2nd) from the analysis since, given a percentage of NAs/observations >30%, even estimates obtained via MICE would have been highly susceptible to bias[3].

3.2 Exploratory and nonparametric analyses

We therefore proceeded to analyze the variables in relation to the overall survival of the patients to provide a first overview of the general situation of the data. As previously said, most of the continuous variables present a skewed distribution, which is also confirmed via graphical approach (Figure 8). Age of the subjects goes from a minimum of 43 years to a maximum of 88 years, with a mean of 68 years and a median of 67 years. 5 people in the [40-50) age group experienced the event of interest, 17 in the [50-60) group, 49 in the [60-70) group, 48 in the [70-80) group and 4 in the [80-90) group. There are a total of 49 censors and 132 events among the patients. As for categorical variables, i.e. stage of the tumour, histology, Gender, World Health Organization Performance Status, number of Lymph nodes attacked and type of protocol chosen for therapy, the events are reported from Table 2 to Table 8. No significative correlations were found among the hematic biomarkers which were included in the dataset. A graphical overview of the correlation levels of the continuous covariates included in the dataset is presented in Figure 9. The first step in every survival analysis is to estimate the overall survival of patients and to examine if the curves related to different levels of the categorical

Stage level	Censors	Events
I	5	15
II	5	20
IIIa	17	32
IIIb	22	65

Table 2: Censors and events for different levels of stage.

Hystotype	Censors	Events
adeno	16	23
NOS	9	57
NCC	24	52

Table 3: Censors and events for different levels of hystology.

Gender	Censors	Events
male	17	41
female	32	91

Table 4: Censors and events for different levels of Gender.

Gender	Censors	Events
male	17	41
female	32	91

Table 5: Censors and events for different levels of Gender.

WHOPS level	Censors	Events
0	21	27
1	23	82
2	5	18
3	0	4
4	0	1

Table 6: Censors and events for different levels of Gender.

Number of lymph nodes attacked	Censors	Events
0	15	30
1	6	22
2	8	26
3	11	14
4	9	40

Table 7: Censors and events for different numbers of lymph nodes attacked.

RTProtcol	Censors	Events
Concurrent RT	35	66
New Protocol August 2005	12	56
Pet Boost	2	10

Table 8: Censors and events for different types of protocol of therapy.

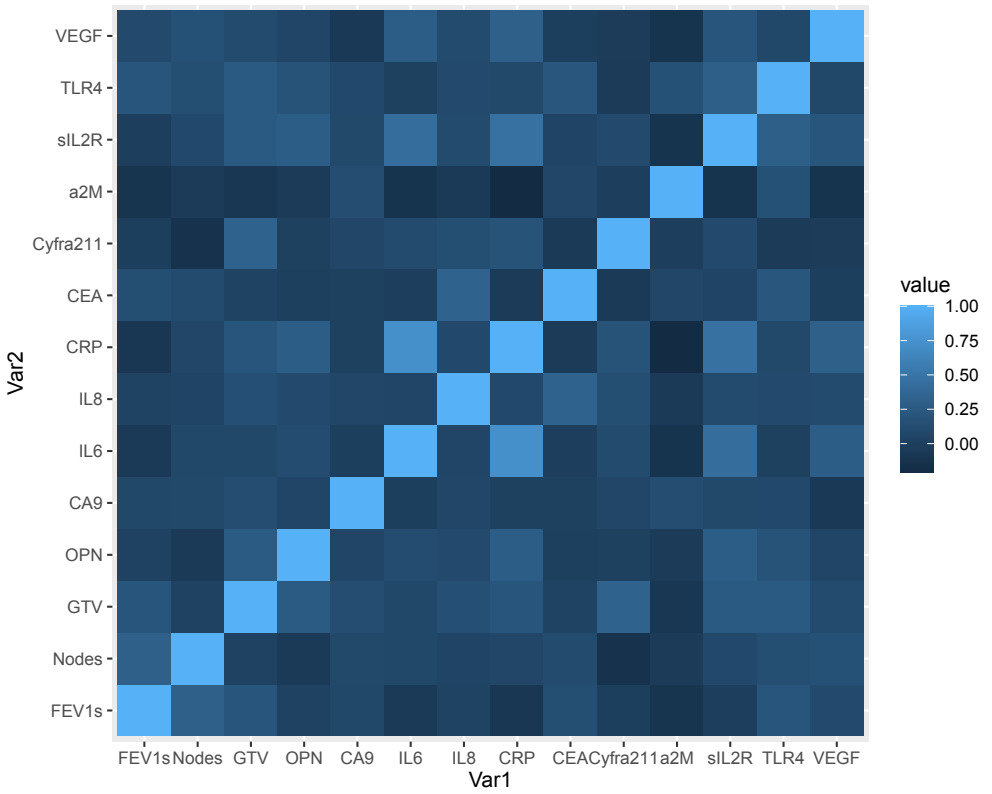

Figure 9: Corrplot related to continuous predictors.

variables significantly differ from each other. If a variable is continuous, it must be categorized into classes. The curves are based and calculated on the Kaplan-Meier estimator. To test if there is a significant difference among the curves for a covariate, a Log-Rank test is performed: if the p-value of the Log-Rank test is significant, then the survival of different levels of the tested variable can not be assumed to be random[16]. Since age is an important predictor of survival in every clinical case, we proceeded to divide it in classes, each with a length of 10 years.

In Figure 10, we illustrate the K.M. survival plots for the overall dataset and for different levels of each categorical covariate, along with the p-value of the Log-Rank test. Each estimated survival curve is comprehensive of the relative 95% confidence interval; these have been removed for stage and age in order to make the plots more readable. The Log-Rank test has proven to be statistically significant at 10% α level for variables related to therapy protocol (p: 0.064) and to histotype (p: 0.056). WHOPS index has proven to be significant at every α level (p: 0.00023); it's important to note that the sample size for its upper level, as it's shown in Table 6, is very low.

Hematic biomarkers were discretized into classes based upon their distribution, also outlier were taken into account when comparing their estimates. As we can see in Figure 11, the blood-biomarkers which resulted as significant to the Log-Rank test, after having been divided in classes, are a2M (p: 0.0012), Cyfra21-1 (p: 0.00037),

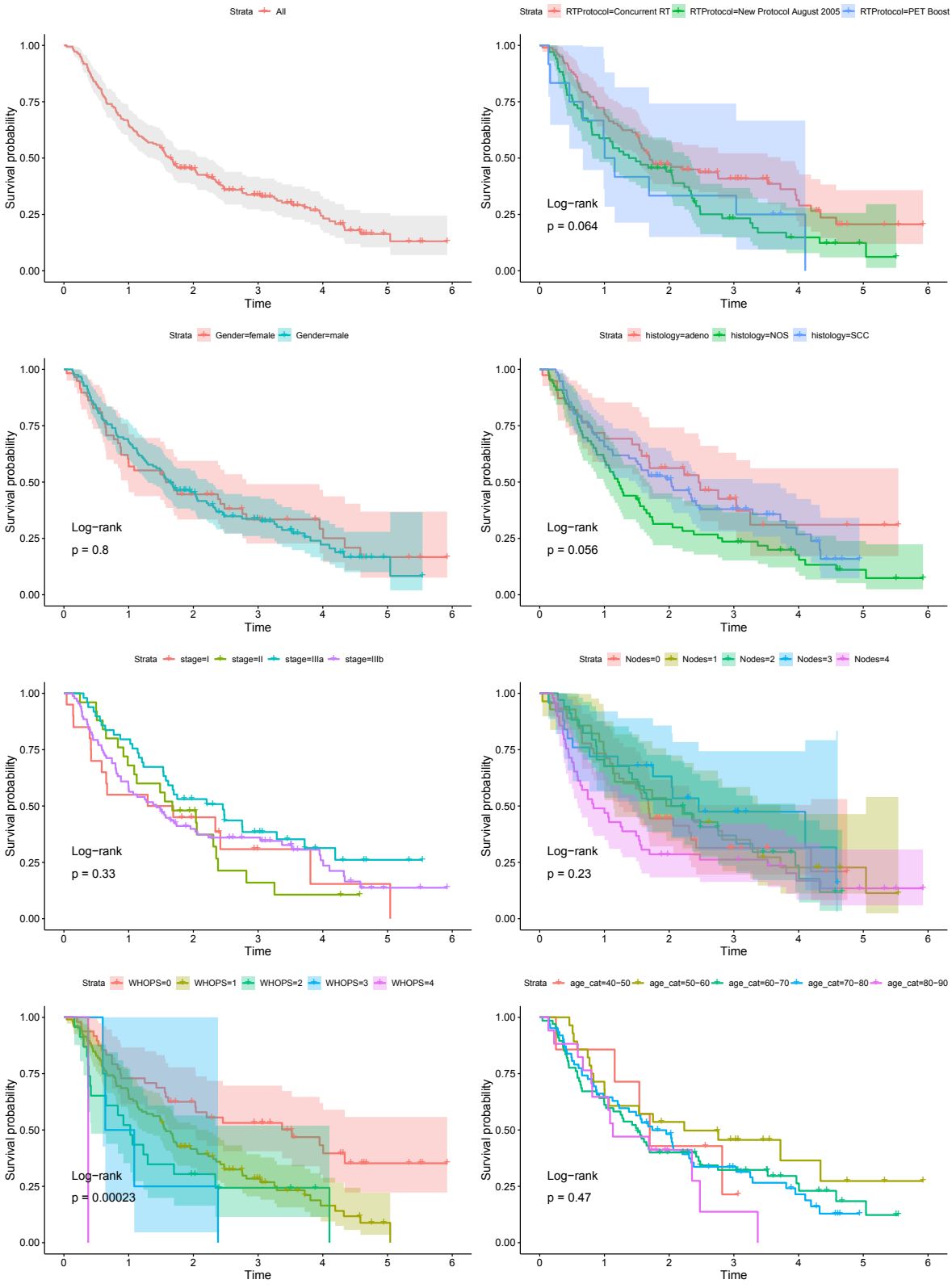

Figure 10: Estimated K.M. survival curves for factors.

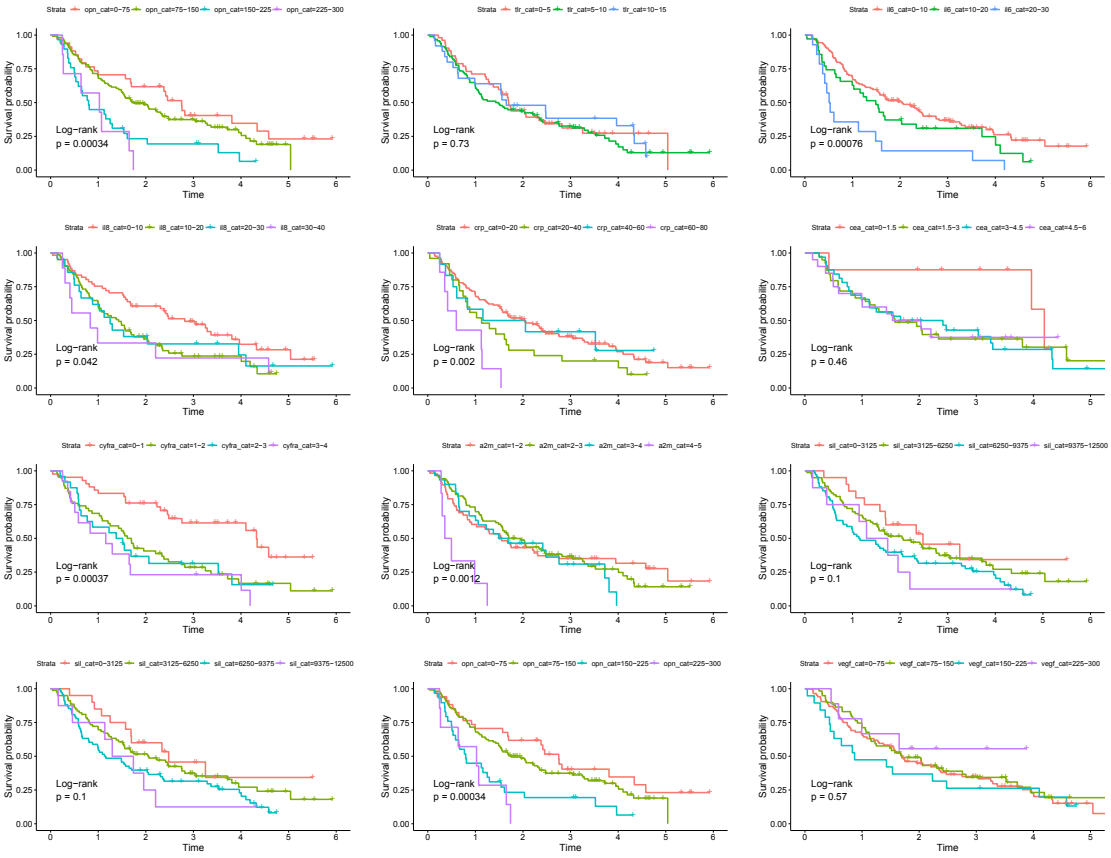

Figure 11: Estimated K.M. survival curves for discretized blood-biomarkers.

CRP (p: 0.002), IL8 (p: 0.042), IL6 (p: 0.0007), OPN (p: 0.00034). Although OPN, a2M and IL6 have a low number of observed events for the upper classes (respectively, 7, 6 and 14). In all 6 cases of these covariates, classes of higher values, i.e. higher blood concentrations of the above biomarkers, always result in worse patient survival expectancy. This result appears to be in line with what was presented in the cited papers[4, 35, 7, 29, 9, 2] and thus confirms the increasing linear relationship between NSCLC mortality and the level of certain immunologic and flogistic predictors.

3.3 Modelling: fitting, selection and comparisons

The starting point of this dissertation, in terms of modelling, was represented by application of the Cox semi-parametric proportional hazards model to the data.
The first model, fitted by including all covariates included in the dataset except for TotalDose(2nd), was defined as it follows:

$$h_i(t|Z) = h_0(t)e^{Z_1\beta_1+Z_2\beta_2+...+Z_{21}\beta_{21}} \tag{66}$$

where:
Z_1: age, Z_2: stage, Z_3: histology, Z_4: Gender, Z_5: WHOPS, Z_6: Nodes, Z_7: RTProtocol, Z_8: TotalDose1st, Z_9: GTV, Z_{10}: OPN, Z_{11}: CA9, Z_{12}: IL6, Z_{13}: IL8, Z_{14}: CRP, Z_{15}: CEA, Z_{16}: Cyfra211, Z_{17}: a2M, Z_{18}: sIL2R, Z_{19}: TLR4, Z_{20}: VEGF, Z_{21}: FEV1s.
Selection of the variables was then performed via AIC, taking also account for the biomarkers related to cancer parameters, which led to the following final model:

$$h_i(t|Z) = h_0(t)e^{Z_1\beta_1+Z_2\beta_2+...+Z_{11}\beta_{11}} \tag{67}$$

where:
Z_1: Gender, Z_2: CEA, Z_3: histology, Z_4: WHOPS, Z_5: Nodes, Z_6: RTProtocol, Z_7: GTV, Z_8: OPN, Z_9: a2M, Z_{10}: sIL2R, Z_{11}: TLR4.
The summary of estimates related to this model is presented in Table 9. The exponential of the estimated coefficient, i.e. the H.R. (hazard ratio) for male subjects, considering female ones as a baseline, doesn't seem to indicate that there's difference in the mortality for this covariate (H.R.: 0.995, C.I.: [0.631, 1.569], p: 0.982). The same can be asserted for CEA and for biomarkers related to GTV, OPN, a2M, sIL2R, TLR4. OPN, a2M and sIL2R are significant at every level of α. Subjects with a NOS histotype have an addition of 25% of the risk (C.I.:[0.729, 2.162], p: 0.411) while those with SCC have an overall decrease of the mortality rate by 18% (C.I.:[0.729, 2.162], p: 0.509) with respect to adenocarcinoma subtype patients. Estimates for the risk of different levels of WHOPS are coherent with the representative nature of said index, although an extremely wide interval for the upper level (H.R.: 15.665, (C.I.:[1.941, 126.450], p: 0.009) should be carefully reconsidered upon increase of the sample size

Covariate	Coef.	exp(Coef.)	SE(Coef.)	z	P(>|z|)	Lower 0.95	Upper 0.95
Gender male	-0.0052216	0.9947921	0.2324791	-0.022	0.982081	0.6307	1.569
CEA	-0.0009898	0.9990107	0.0007496	-1.320	0.186685	0.9975	1.000
histology NOS	0.2277218	1.2557359	0.2771691	0.822	0.411305	0.7294	2.162
histology SCC	-0.1863336	0.8299967	0.2824956	-0.660	0.509512	0.7294	2.162
WHOPS1	0.6487922	1.9132286	0.2393567	2.711	0.006717**	1.1968	3.058
WHOPS2	0.8186976	2.2675446	0.3592306	2.279	0.022665*	1.1214	4.585
WHOPS3	0.9414885	2.5637948	0.6173533	1.525	0.127249	0.7645	8.598
WHOPS4	2.7514329	15.6650617	1.0655363	2.582	0.009817**	1.9406	126.450
Nodes	0.2079295	1.2311264	0.0689194	3.017	0.002553**	1.0756	1.409
Protocol Aug. 2005	0.4297267	1.5368374	0.2217192	1.938	0.052604	0.9952	2.373
Protocol PET Boost	-0.2276432	0.7964084	0.3935071	-0.578	0.562928	0.3683	1.722
GTV	0.0011168	1.0011174	0.0006214	1.797	0.072321	0.9999	1.002
OPN	0.0024168	1.0024197	0.0006663	3.627	0.000287***	1.0011	1.004
a2M	0.3375303	1.4014822	0.1222630	2.761	0.005768**	1.1029	1.781
sIL2R	0.0001355	1.0001355	0.0000338	4.010	6.08e-05***	1.0001	1.000
TLR4	-0.0437253	0.9572169	0.0259773	-1.683	0.092334	0.9097	1.007

Table 9: Summary of the Cox model with covariates selected via AIC.

Covariate	Chisq.	df	P(>|z|)
Gender	0.542	1	0.461
CEA	4.055	1	0.044*
histology	0.980	2	0.613
WHOPS	3.755	4	0.440
Nodes	5.802	1	0.016**
RTProtocol	0.184	2	0.912
GTV	0.132	1	0.717
OPN	0.241	1	0.623
a2M	1.039	1	0.308
sIL2R	0.298	1	0.585
TLR4	4.542	1	0.033*
Global	26.355	16	0.037*

Table 10: Test for proportional hazards for the Cox model with covariates selected via AIC.

(only one subject was observed for this level). As for lymph nodes which showed to be attacked by the tumor under PET examination, an increase of one unit, meaning one more lymph node involved in the carcinogenic process, corresponds to an increase in the risk by 23% (C.I.:[1.075, 1.409], p: 0.002). Therapy related to the August 2005 Protocol has proven to be, under this model, less efficient (H.R.: 1.5368, C.I.:[0.995, 2.373], p: 0.052) than the one with PET Boost (H.R.: 0.7964, C.I.:[0.358, 1.722], p: 0.563) when compared to the baseline (Concurrent RT).

We therefore proceeded to evaluate the adequacy of the model under the null hypothesis of proportional hazards with the relative proportional hazards test. Different transformations for the survival times (i.e.: logarithmic, rank, identity, Kaplan-Meier's) did not lead to any substantial variation in the results. Results for this test are shown in Table 10.

The model appears to be overall significant in terms of violating the null hypothesis (Chisq.: 26.355, df: 16, p: 0.037) at level of α 5%. Single regressors which have proven

to violate the null hypothesis are CEA (Chisq.: 4.055, df: 1, p: 0.04), Nodes (Chisq.: 5.802, df: 1, p: 0.016) and TLR-4 (Chisq.: 4.542, df: 1, p: 0.03).
A first comparison among the overall estimated survival curves for the data and under the Cox model is shown in Figure 12. Survival appears to be overestimated under the Cox model with respect to the one obtained via application of the Kaplan-Meier estimator to the data. Diagnostics for this model were particularly driven to the aim of

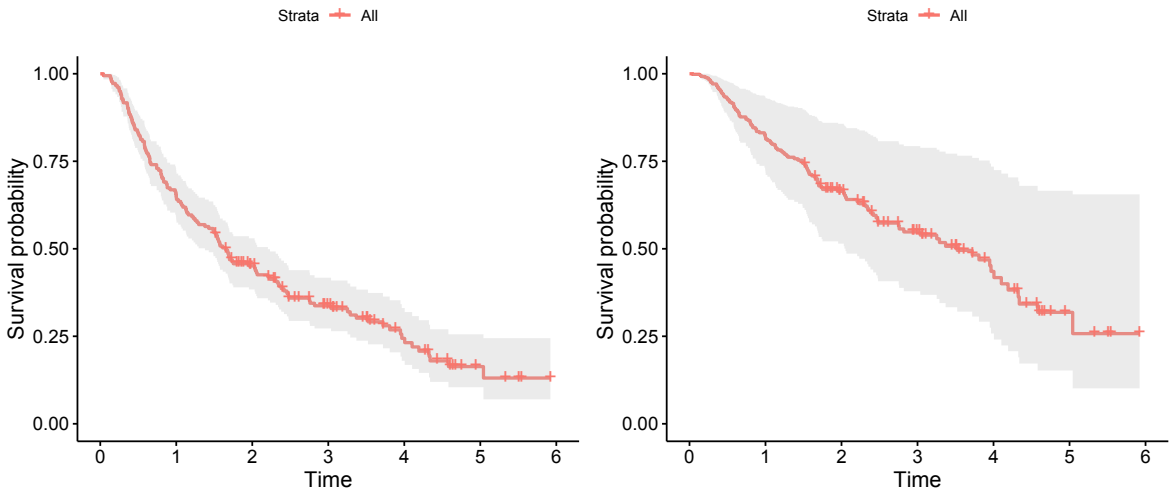

Figure 12: Comparison between estimated survival curves: nonparametric Kaplan-Meier estimator (sx) v. Cox model (dx).

verifying the hypothesis of proportional hazards, taking account of these evidences. In Figure 13 we illustrate the Schoenfeld residuals plot for covariates which seem to violate the null hypothesis, according to the previous test. Score processes with 50 unweighted simulated processes under the null hypothesis of the Cox model were done in order to further investigate the behaviour on the risk over time for these three covariates. Plots are shown in Figure 14: all three of them seem to indicate that a deviation from the null hypothesis of proportional hazards is indeed present. Because of this, it was concluded that the standard Cox semi-parametric proportional hazards model was not fully able to represent the effects of the covariates over time and, because of this, its time-varying version was fitted. Following the approach of Martinussen & Scheike[22], it's possible to include all covariates in the extended Cox model as time-varying and then test for the linearity of their effect on the risk over time. The functional form of the covariates which resulted significantly nonlinear in the output of the model can be then adequately represented upon investigation of the martingale residuals. In his recent article[30], Thurneau suggest that a p-value of how strongly nonlinear the form of a covariate is can be obtained by directly using the **`pspline()`** R command on the term and then by looking at the output of the model. The estimates of the extended Cox model for time-varying effects are presented in Table 11. Upon consecutive re-fittings

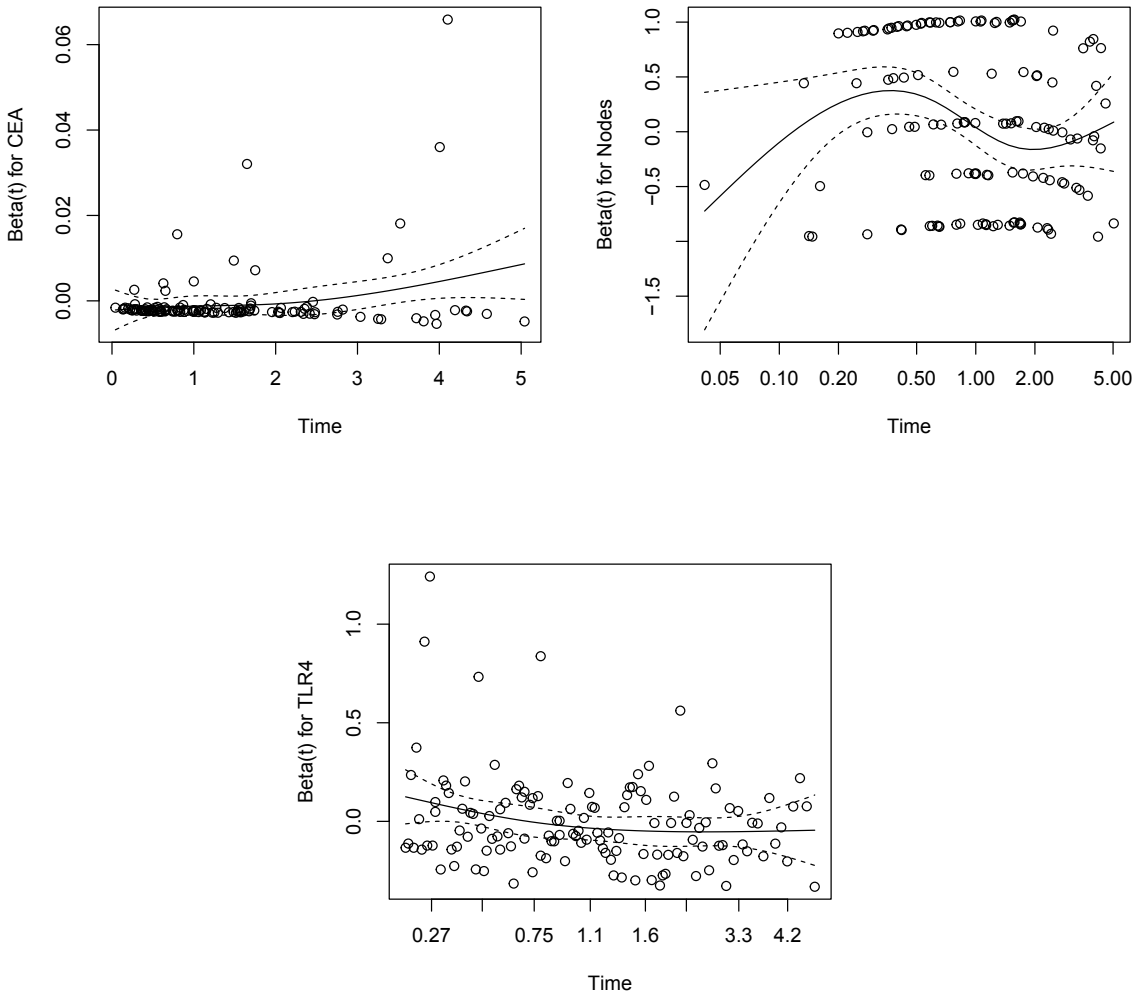

Figure 13: Schoenfeld residuals plots for CEA, Nodes and TLR-4.

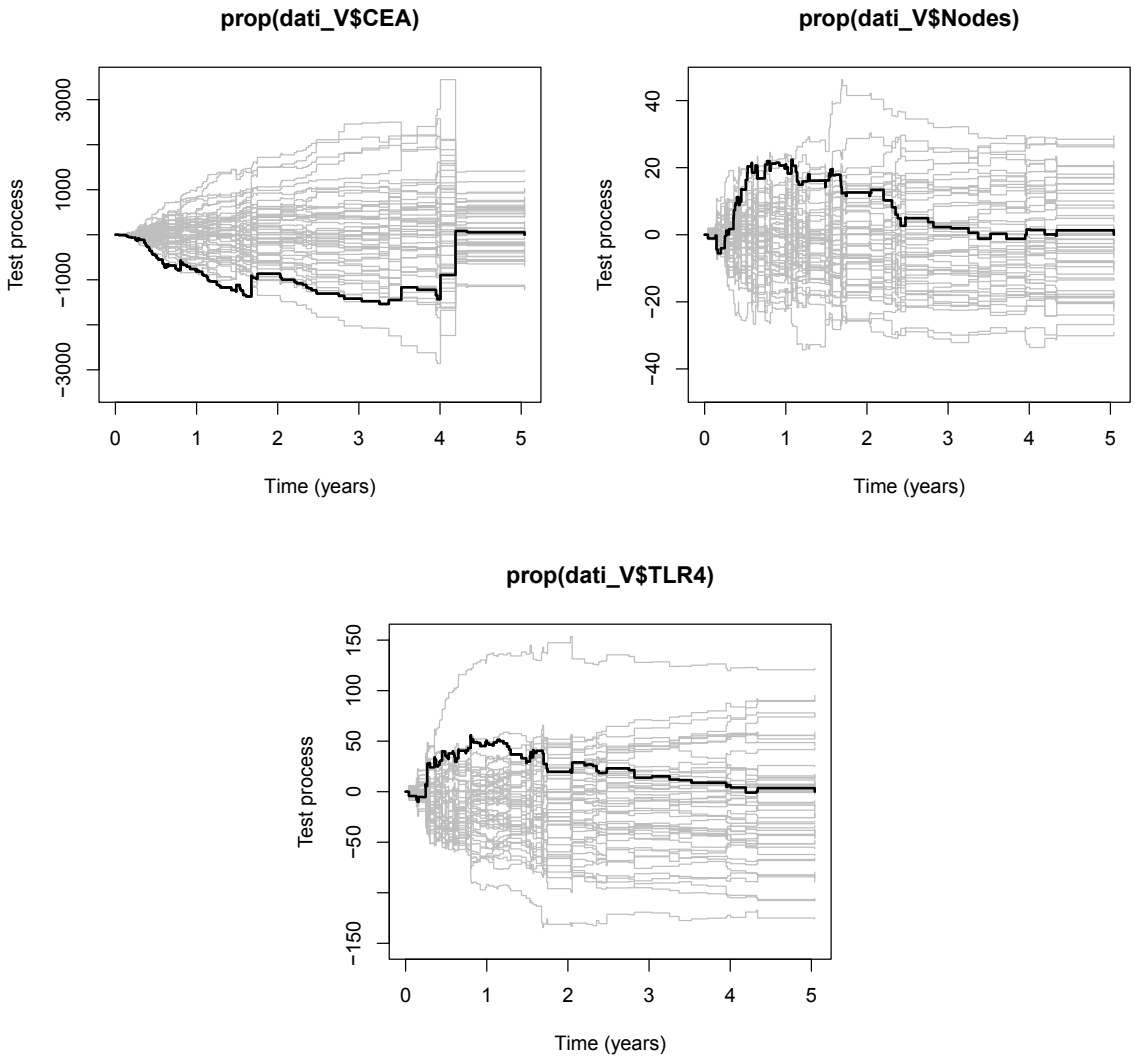

Figure 14: Score processes for with 50 unweighted simulated processes under the null hypothesis for CEA, Nodes and TLR-4.

Covariate	exp(Coef.)	SE(Coef.)	Chisq.	df	P(>|z|)	Lw0.95	Up0.95
Gender male	1.02	0.24	0.01	1.00	0.9200	0.63	1.69
CEA	0.99	0.0007	4.00	1.00	0.0450*	0.997	1.000
histologyNOS	0.99	0.285	0.00	1.00	0.9800	0.567	1.733
histologySCC	0.63	0.286	2.49	1.00	0.1100	0.363	1.11
WHOPS1	2.16	0.245	10.18	1.00	0.0014***	1.348	3.488
WHOPS2	3.35	0.376	10.30	1.00	0.0013***	1.60	7.01
WHOPS3	3.81	0.632	4.50	1.00	0.0340*	1.11	13.17
WHOPS4	33.17	1.089	10.34	1.00	0.00133***	3.93	280.26
Nodes	1.22	0.072	7.40	1.00	0.0065***	1.056	1.358
Protocol Aug. 2005	1.54	0.224	3.82	1.00	0.0510	0.998	2.388
Protocol PET Boost	0.78	0.412	0.34	1.00	0.5600	0.349	1.763
Pspline(GTV, linear)	1.00	0.0004	1.846	1.00	0.08	0.99	1.004
Pspline(GTV, nonlinear)	11.75	3.22	0.34	1.00	0.001***	1.8	74.17
OPN	1.00	0.0007	6.93	1.00	0.0085***	1.00	1.03
a2M	1.27	0.128	3.83	1.00	0.0500	0.99	1.61
sIL2R	1.00	3.5e-05	9.20	1.00	0.0024***	1.00	1.00
TLR-4	0.97	0.03	0.87	1.00	0.3500	0.92	1.08

Table 11: Summary of the extended Cox model for time-varying effects after removal of the non-significant nonlinear parts of the regression terms.

of the initial model with all covariates being time-varying, the only covariate which exhibited a significant nonlinear trend over time was GTV. As discussed in section 1.2.3, a penalized spline approach was implemented to best represent the functional form of this covariate in the final model. The number of degrees of freedom selected as an ideal representation of its functional form was five. In Figure 15 we can observe martingale residuals for GTV and how they can not be properly described by a straight line; we then proceeded to interpolate its functional form with a penalized spline term as previously explained, the graphical representation of the p-spline interpolating GTV is shown in Figure 16. Estimates computed under the extended Cox model for time-varying effects show a decrease of the effect of a2M (H.R.: 1.27, C.I.:[0.99, 1.61], p: 0.05) if compared with the previous model. Estimated risk for Gender does not seem to vary much between one model and the other (0.03% difference for the exponential of the estimated coefficient). All three levels of WHOPS, compared to the baseline (WHOPS: 0), as can be seen by looking at Table 9 and Table 11, differ: the model with time-varying effects appears to overall provide higher estimates. No substantial change can be observed for covariates related to lymph nodes and CEA.

What certainly looks evident is an overestimation of risk for GTV, whose functional form was interpolated with a p-spline with 5 degrees of freedom: an increase of one unit for this hematic biomarker leads us to assume an increase of almost twelve times the risk, the upper limit of the confidence interval being also equal to almost 75. This downward is most likely due to the fact that the estimation strongly relies on the number of degrees of freedom selected for the penalized spline term of GTV.

This leads us to the necessity of validating the accuracy of the extended Cox model

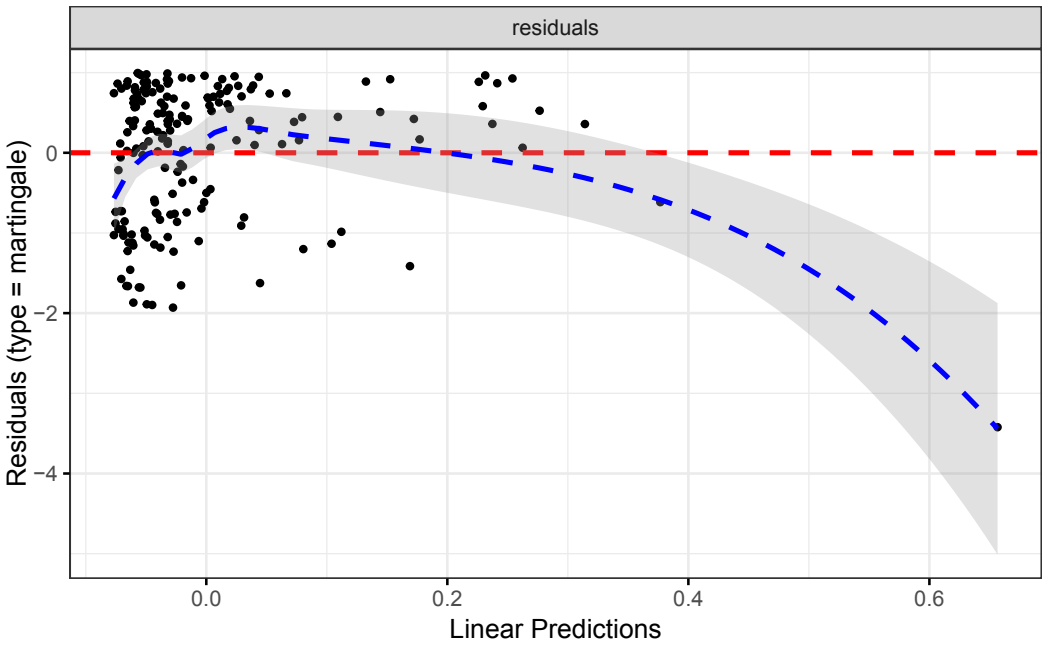

Figure 15: Functional form evaluation of GTV by inspection of its martingale residuals.

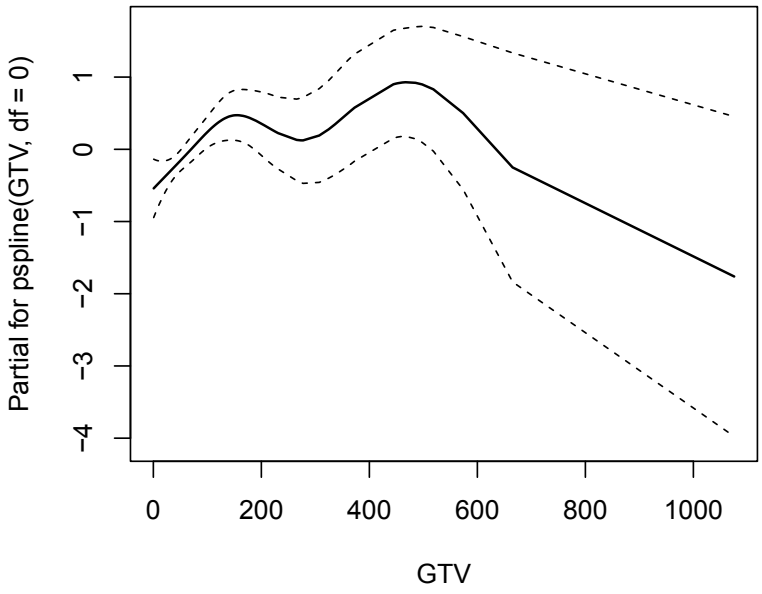

Figure 16: Penalized spline term for GTV, based on the previous martingale residuals inspection. Degrees of freedom are equal to five.

for time-varying effects even after having considered p-spline transformations for those variables which did not exhibit costant effect over time. The multiplicative approach alone might not be well suited to adequately describe our data, or it might be appropriate to, at least, try a different strategy. We therefore proceeded to fit the Aalen additive hazards regression model, McKeague & Sasieni additive hazards model and Cox-Aalen multiplicative-additive hazards model.
We start by considering the following Aalen additive-hazards model:

$$h(t|Z) = Z^T \alpha(t) \tag{68}$$

in which Z includes the regressors related to GTV, age, stage, histology, Gender, Nodes, RTProtocol, CEA, Cyfra21-1, a2M, TLR-4 and VEGF; other covariates had to be omitted because of convergence issues of the model which could not be overcome even by allowing the number of iterations to increase; this issue is related to the R *timereg* package, whose models, we noticed, can not always reach convergence when several covariates or when spline terms are included. The approach for fitting this model is similar to the one illustrated for the extended Cox model for time-varying effects: all covariates are initially included in the model as time-varying. Then, a Supremum test of significance and a test for time-constant effects are performed. To ensure best flexibility, the model is then re-fitted with the covariates which were resulted as not significant from the test, being inserted as time-constant, and the other ones as time-varying: this results in the semi-parametric additive hazards regression model of McKeague & Sasieni[22].
We first fit the Aalen model and look directly at the output, which is shown in Table 12, to see which covariates are significant for the Supremum test and which are significant for the time-constant effects test. Using the Supremum test, we see that the number of lymph nodes, therapy protocol and CEA are significant at 5% α level.
Looking at the results in Table 12, we now start to simplify the model by a number of successive tests with the purpose of reducing the number of nonparametric components, since stageII, Nodes, Protocol, CRP and CEA seem to have a significant time-varying effect on the risk, while the other covariates don't.
This means we're fitting the previously mentioned semi-parametric additive hazards model of McKeague & Sasieni, which is defined as it follows:

$$h(t) = Z^T \alpha(t) + X^T(\gamma) \tag{69}$$

The set of regressors X, whose effects are fixed in time, coherently with the aforementioned results, are GTV, age, histology, Gender, Cyfra21-1, a2M, TLR-4 and VEGF.
P-values of the second test for time-invarying effects is shown in Table 13. To further explore the behaviour over time on the risk of the covariates which were significant

Covariate	Sup. test of significance	P(>|z|) (Sup. test of significance)	P(>|z|) (Time-constant effects)
(Intercept)	1.12	0.815	0.669
WHOPS1	2.01	0.215	0.535
WHOPS2	1.58	0.473	0.17
WHOPS3	1.75	0.232	0.185
WHOPS4	25.7	0.197	0.112
GTV	1.93	0.321	0.128
age	2.13	0.199	0.484
stageII	2.48	0.069	0.023*
stageIIIa	1.71	0.344	0.237
stageIIIb	1.55	0.443	0.299
histologyNOS	1.19	0.827	0.590
histologySCC	1.87	0.301	0.465
Gender male	1.98	0.291	0.133
Nodes	3.15	0.025*	0.004**
Prot.Aug.2005	2.99	0.017*	0.008**
Prot.PET Boost	1.85	0.285	0.08
CRP	2.55	0.073	0.023*
CEA	3.06	0.027*	0.013*
Cyfra21-1	1.52	0.680	0.36
a2M	2.08	0.276	0.611
TLR-4	1.74	0.460	0.383
VEGF	1.43	0.673	0.255

Table 12: P-values for the Supremum test of significance and for the test for time-invarying effects of the initial Aalen model.

Covariate	P(>|z|) (Time-constant effects)
(Intercept)	0.059
stageII	0.255
stageIIIa	0.308
stageIIIb	0.293
Nodes	0.043*
Prot. Aug. 2005	0.425
Prot. PET Boost	0.553
CRP	0.682
CEA	0.004**

Table 13: P-values of the second test for time-invarying effects for the reduced model of McKeague & Sasieni.

for the second time-constant effects test, we plotted the cumulative coefficients of the model: if the effects are not time-varying, they should approximately be well-described by a straight line[22]. The only two covariates which did not satisfied this indication were Nodes and CEA, as it's shown in Figure 17. This is consistent with the fact that Nodes and CEA were the only two covariates which, again, resulted as significant in the output of the test for McKeague & Sasieni's model, we therefore conclude that the effects of these two covariates on the risk are not constant over time, and we fit the final model of McKeague & Sasieni whose estimates for parametric terms are illustrated in Table 14. Given this, we fit the final model whose output is illustrated in Table

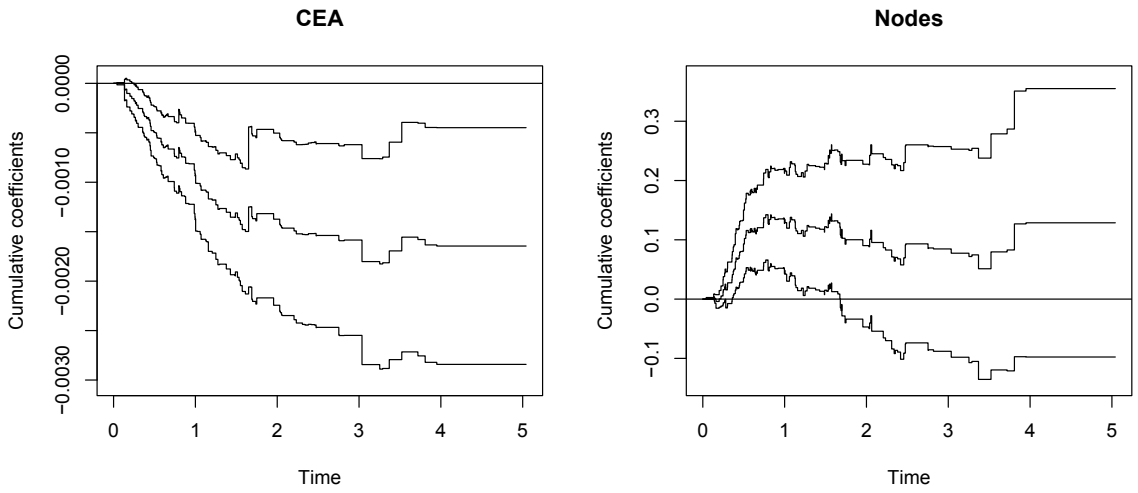

Figure 17: Cumulative coefficients plots for CEA and Nodes with 95% pointwise confidence intervals: both are not well described by a straight line.

14. Looking at the estimates for this model, no covariate seems to show a drastically increase in terms of effect. Histotype NOS has an increased estimated intensity for subjects equal to 0.19 (C.I.:[-0.02, 0.418], p: 0.04) and subjects with therapy protocol August 2005 have an increased estimated intensity of 0.179 (C.I.:[-0.0017, 0.375], p: 0.082). What we can note is that the impact of these regressors over the risk differs a lot from the previous Cox models, thus indicating how different modelling approaches may result in different conclusion of the results. The increasing effect on the risk for a higher score of the WHOPS index is coherent with the estimates of the previous Cox model (and with the Cox-Aalen model which follows). Nevertheless, one problem which arises from this model, as discussed in section 2.2.2, is that the estimates for some coefficients appear to be negative or very close to zero. Given this, we proceed with possibly the best compromise in terms of flexibility so far which is represented by the multiplicative-additive hazards model of Cox-Aalen.

Covariate	Coef.	SE	Robust SE	z	P-val	low2.5%	up97.5%
const(WHOPS)1	0.219352	0.0675	0.0638	3.43	0.000609	0.0867	0.351
const(WHOPS)2	0.360352	0.1410	0.1550	2.33	0.0199	0.0836	0.636
const(WHOPS)3	0.616754	0.4280	0.2640	2.34	0.0193	-0.2230	1.450
const(WHOPS)4	2.560552	2.67	0.0785	32.50	0.0523	-2.67	7.790
const(GTV)	0.000344	0.000261	0.000293	1.170	0.24100	-0.000168	0.000856
const(age)	0.000799	0.005060	0.004170	0.192	0.84800	-0.009120	0.010700
const(stage)II	-0.066700	0.155000	0.156000	-0.427	0.66900	-0.370000	0.237000
const(stage)IIIa	-0.117000	0.157000	0.167000	-0.701	0.48400	-0.425000	0.191000
const(stage)IIIb	-0.083600	0.153000	0.163000	-0.513	0.60800	-0.383000	0.216000
const(hist.)NOS	0.195000	0.114000	0.096500	2.020	0.04370	-0.028400	0.418000
const(hist.)SCC	-0.040700	0.097500	0.090300	-0.451	0.65200	-0.232000	0.150000
const(Gender)male	-0.030200	0.103000	0.090400	-0.334	0.73800	-0.232000	0.172000
const(Prot.Aug.2005)	0.179000	0.100000	0.103000	1.740	0.08200	-0.017000	0.375000
const(Prot.PET Boost)	0.109000	0.196000	0.186000	0.586	0.55800	-0.275000	0.493000
const(CRP)	0.002200	0.001180	0.001210	1.810	0.06960	-0.000113	0.004510
const(Cyfra211)	0.009340	0.005160	0.002910	3.210	0.00131	-0.000773	0.019500
const(a2M)	0.106000	0.062700	0.058800	1.810	0.07080	-0.016900	0.229000
const(TLR4)	-0.005860	0.010300	0.009450	-0.620	0.53500	-0.026000	0.014300
const(VEGF)	0.000186	0.000394	0.000462	0.402	0.68800	- 0.000586	0.000958

Table 14: Estimates of parametric terms of the reduced model of McKeague & Sasieni.

The model is fitted as it follows:

$$h(t) = (X^T(t)\alpha(t))e^{Z^T\beta} \tag{70}$$

Since we learned from the previous models that the only two covariates which exhibit time-varying effects are Nodes and CEA, these last two covariates are incorporated in the flexible additive part of the model, while all other previous covariates are included in the multiplicative part. The summary of the parametric components of the Cox-Aalen model is illustrated in Table 15 along with the p-values of the proportionality hazards test, while the one related to the usual tests performed on the components of the additive part is illustrated in Table 16. First, we note that both Nodes and CEA are still significant for the test of time-invariant effects (respectively, p.: 0.006 and p.: 0.002), while every other covariate included in the multiplicative part of the model does satisfy the proportional hazards hypothesis. Under these premises, the Cox-Aalen model would seem to fit the data reasonably well.

Looking at the estimates of the parametric part, the only covariates which result significant at 5% α level for this model are the three hematic biomarkers a2M, CRP and Cyfra21-1. All four levels of WHOPS, compared to the baseline WHOPS0 (always taking into account for the fact that WHOPS4 has a really low number of observed events) are significant at 5% α level; considering the exponential of the estimated coefficients, a subject with a WHOPS score equal to 1 has an increased risk of 69% (C.I.: [1.05, 2.77], p: 0.025), 2 has an increase by 174% (C.I.:[1.28, 5.87], p: 0.006), and 3 by 278% (C.I.: [0.99, 14.43], p: 0.04). Although a higher number of subjects would be useful to deter-

Covariate	Coef.	SE	z	P-val	low2.5%	up97.5%	Prop.Test(z)	Prop.Test(P-val)
WHOPS1	0.53	0.24	2.24	0.025*	0.05	1.02	5.27	0.198
WHOPS2	1.01	0.39	2.7	0.006**	0.25	1.77	4.64	0.052
WHOPS3	1.33	0.68	2.02	0.04*	-0.009	2.67	1.16	0.5
WHOPS4	2.97	1.07	6.77	1.33e-11	0.873	5.07	0.879	0.3
GTV	0.0009	0.0006	1.44	0.14	-0.0002	0.002	874.0	0.818
age	0.06	0.013	0.53	0.59	-0.01	0.031	97.9	0.22
stageII	-0.217	0.355	-0.59	0.54	-0.91	0.5	3.8	0.164
stageIIIa	-0.32	0.38	0.79	0.43	-1.06	0.42	5.42	0.11
stageIIIb	-0.152	0.353	-0.401	0.69	-0.84	0.54	4.48	0.298
histologyNOS	0.478	0.289	1.91	0.057	-0.088	1.04	5.06	0.274
histologySCC	-0.045	0.294	-0.163	0.871	-0.621	0.531	3.21	0.766
Gender Male	-0.117	0.266	-0.496	0.62000	-0.63800	0.4040	5.68	0.104
Prot.Aug.2005	0.474	0.247	1.82	0.069	-0.01	0.95	5.94	0.106
Prot.PET Boost	0.372	0.429	0.921	0.357	-0.469	1.21	1.70	0.688
CRP	0.005	0.002	2.69	0.007**	9e-4	0.009	288.00	0.646
Cyfra21-1	0.013	0.005	2.75	0.006**	0.002	0.025	187.00	0.128
a2M	0.247	0.131	1.92	0.05*	-0.01	0.504	9.36	0.132
TLR-4	-0.018	0.027	-0.677	0.5	-0.07	0.036	53.30	0.160
VEGF	0.0006	0.001	0.502	0.62	-0.002	0.003	930.00	0.28

Table 15: Estimates of parametric terms (multiplicative part) of the Cox-Aalen model; each covariate is also tested for proportionality of the hazards.

Covariate	Time-invariant effects test(z)	Time-invariant effects test(P-value)
Intercept	1.25	0.44
Nodes	3.47	0.006**
CEA	4.21	0.002**

Table 16: Test for time-invariant effects for the components of the additive part of the Cox-Aalen model.

mine it, it's safe to assume that a higher WHOPS score does indeed result in a worse outcome for the survival of patients with NSCLC. This is in line with the result of the Cox for time-varying effects and with the model of McKeague & Sasieni. In Figure 18 the cumulative coefficients plots for Nodes and CEA are illustrated, both now appear to be able to be described by a straight line much better than in the previous model although they still resulted as significant for the test for time-varying effects. In Figure 19, we illustrate the score processes for the multiplicative components of the Cox-Aalen model with 50 realizations. Comparison among the survival curves estimated under

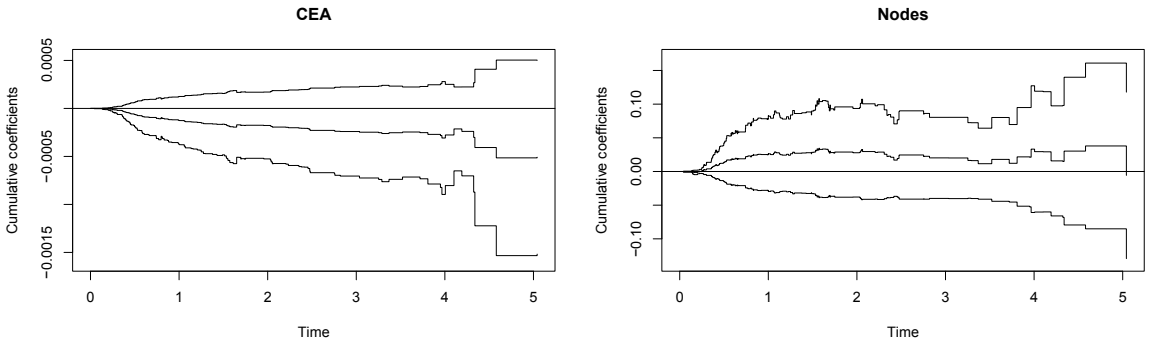

Figure 18: Cumulative coefficients plots for CEA and Nodes (Cox-Aalen model) with 95% pointwise confidence intervals.

the Cox model, the Cox model for time-varying effects, the McKeague & Sasieni's model and the Cox-Aalen model has met the necessity of including only a certain part of the initial set of covariates of the dataset, either because some of them were selected via AIC at the preliminary stage of the analyses, or because some of them, due to limitations of the R *timereg*[22] package that implements the Aalen, McKeague & Sasieni, and Cox-Aalen models, do have convergence issues when the number of regressors exceeds a certain threshold. Upon these considerations, the regressors included when we compared the five different curves were: Gender, RTProtocol, WHOPS, GTV, stage, histology, Cyfra21-1, a2M and TLR-4. That being said, we illustrate the comparison of the estimated survival curves under the three models in Figures 20 and 21. What can be observed, in terms of predictive abilities of the three models, is first of all how the Cox model largely overestimates survival compared with the reduced McKeague & Sasieni reduced model, and moderately compared with the Cox-Aalen model. To quantify this comparison we note that the median survival time for the Cox model is about 3 years and 11 months, for the McKeague & Sasieni model it is about 1 year and 2 months, and for the Cox-Aalen model it is about 2 years. The gap between the predicted survival estimate under the Cox model should be viewed with great caution, as we can see that the confidence interval increases greatly as time advances, particularly at the upper extreme; this reflects the instability in the production of the estimates that had also been seen with regard to the handling of GTV using a penalized spline. On the

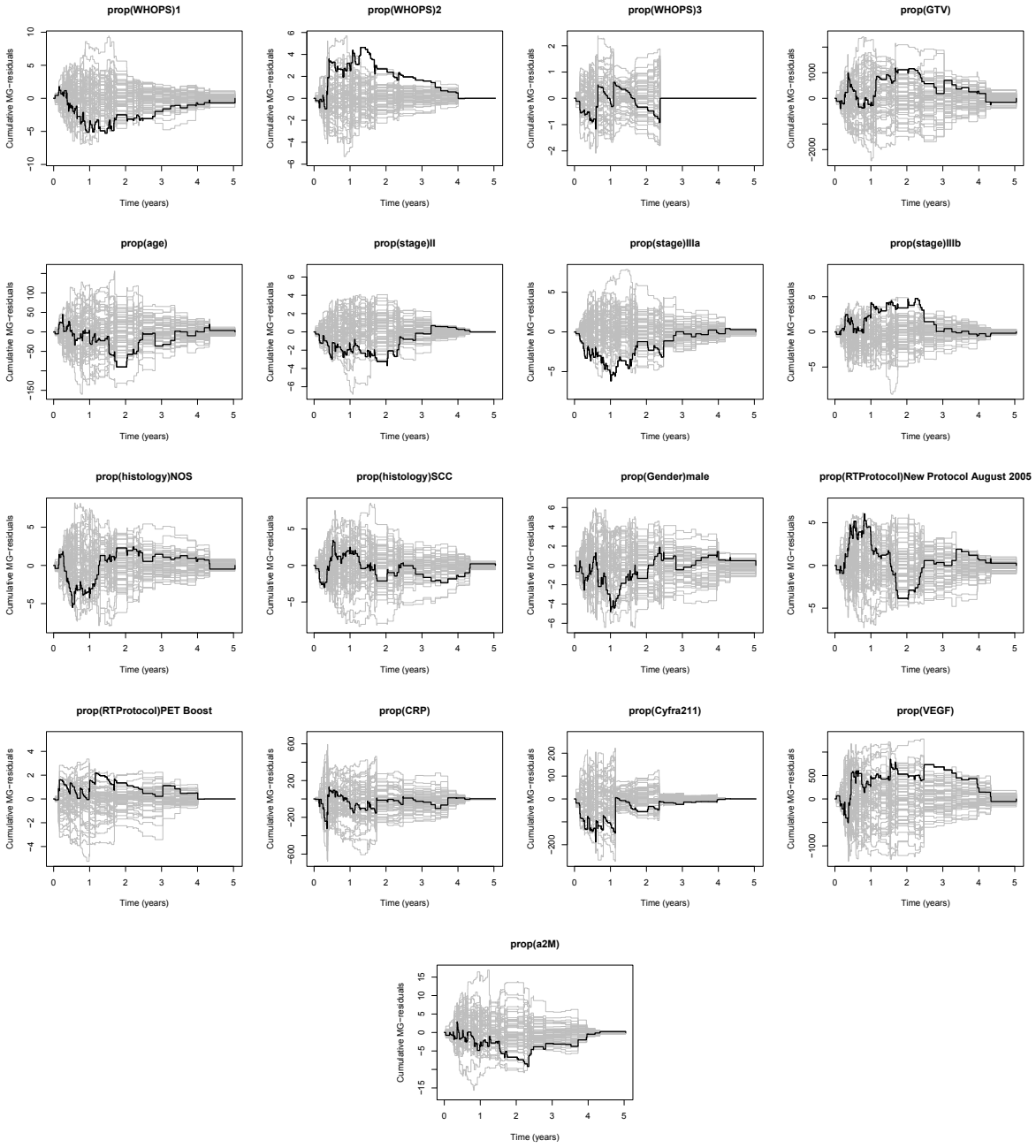

Figure 19: Score processes for multiplicative part of Cox-Aalen model with 50 random realizations under the model.

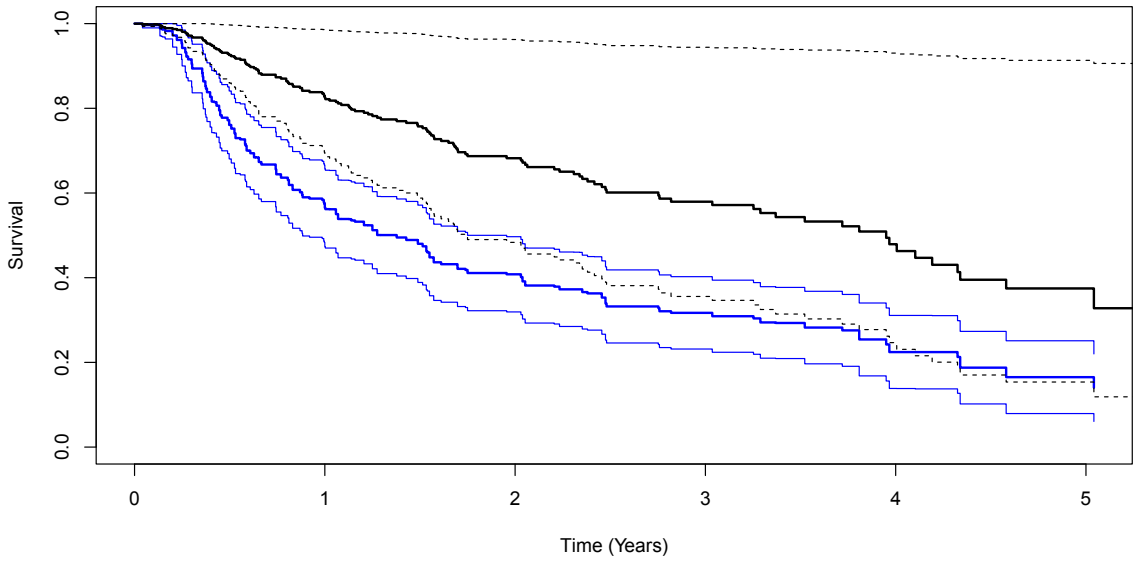

Figure 20: Comparison among estimated survival curves under the flexible model of McKeague & Sasieni (in blue) and the Cox for time-varying effects (in black). 95% confidence bands for the McKeague & Sasieni's model and 95% confidence interval for the Cox model for time-varying effects are also present. Predictions are for patients with age [60-70), stage I, histotype SCC, WHOPS 0 and protocol concurrent RT.

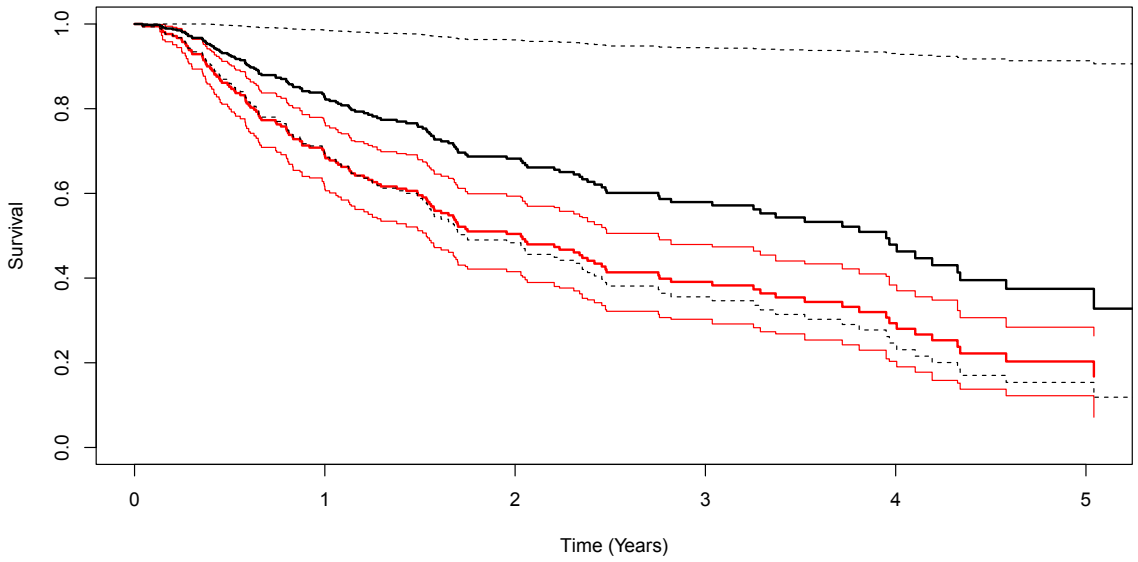

Figure 21: Comparison among estimated survival curves under the flexible model of Cox-Aalen (in red) and the Cox for time-varying effects (in black). 95% confidence bands for the Cox-Aalen model and 95% confidence interval for the Cox model for time-varying effects are also present. Predictions are for patients with age [60-70), stage I, histotype SCC, WHOPS 0 and protocol concurrent RT.

other hand, the reduced McKeague & Sasieni model and the Cox-Aalen model exhibit worse survival performance when compared with the extended version of the previous semi-parametric model; however, they provide a more accurate estimate (looking at the 95% uniform confidence bands) and are more proportional to each other. The median estimated survival time differs for these two models by 8 months; this gap can be considered more plausible than with the Cox model precisely because of the width of the confidence bands. Setting different risk categories combinations has always resulted, obviously in different measures, in an overestimation of the Cox model for time-varying effects in comparison to the models of McKeague & Sasieni and of Cox-Aalen.

Results and conclusions

This thesis aimed to demonstrate how the results in terms of predicted survival can vary significantly when, due to the presence of time-varying effects among the covariates, the Cox model alone is not able to adequately manage the survival modeling. The data underwent an intense data cleaning process due to the presence of several missing values, also using more advanced procedures to set them for analysis, such as MICE. The choice of the type of imputation adopted was based on the number of NAs present, on the nature of the variable and on the symmetry of its distribution.
The main limitations encountered in carrying out these analyses were mostly dictated by software limitations, in particular by the *timereg* package used to fit the flexible models, which is not always able to guarantee convergence when the number of covariates is high or when terms incorporate penalized splines, as in the case of GTV for the time-varying effects Cox model, in order to evaluate its functional form.
Following a brief exploratory analysis which highlighted the distribution of censorships and events for the various predictors, a standard Cox model was initially fitted, whose covariates were selected via AIC; variables inherent to tumor load, number of lymph nodes reported as positive by PET and blood-biomarker TLR-4 violated the proportional hazards assumption of the model.
Estimates of blood-biomarkers confirmed the findings of the aforementioned studies which explained that higher concentrations in the circulatory system of a2M, Cyfra21-1, CRP, IL8, IL6, and OPN correspond to a worse survival expectancy for patients with NSCLC.
For the time-varying effects Cox model, the approach that was chosen aimed to directly describe the functional form of the covariates that were significant in a test based on their nonlinearity on the risk over time, this was the case for GTV. A poor fitting of this last model, particularly in relation to the variable whose functional form has been interpolated via p-spline, has led to the usage of several more flexible models, in which a compromise between parametric and non-parametric components is performed, to take into account those variables that demonstrate effects on the risk that vary over time.
The Aalen model highlighted how the covariates which exhibited this behavior were Nodes and CEA: this is consistent with the nature of the variables, as we can expect that the number of lymph nodes attacked by the tumor and the tumor load both in-

crease over the course of time and consequently have a more incisive influence on the survival outcome.
Having established this, these two covariates have been included in the part that can vary over time, while all the others have been incorporated in the parametric part. The semi-parametric approach within the family of additive models results in what is called the semi-parametric McKeague and Sasieni's additive risks model. The biggest limitation of additive risks models is that estimates can often be negative or very close to zero. To deal with this eventuality, which then effectively occurred, a Cox-Aalen model was fitted: said model incorporated the two covariates previously found to be significant with the test for time-constant effects: these were then incorporated in the additive part, while all the others in the multiplicative part. The exponential operator ensures that non-negative estimates for the parametric components are produced.
While estimates for blood-biomarkers did not show particular increases or decreases of the effect on the risk, a variable that was always significant in all models was the WHOPS indicator: higher levels of this indicator always corresponded to worse outcomes in terms of survival for patients, regardless of the type of therapy adopted or the stage of the tumor. What can be deduced from this, and which is in line with the nature of the indicator, is that a higher score and consequently an unfavorable health situation of the patient with NSCLC results in worse survival expectations for this condition.
Compared to the *concurrent radiotherapy* type of therapy, all models predicted a worse outcome for patients with the *August 2005* type of therapy than for those with the *PET Boost* type of therapy.
Different stages of advancement of the neoplasm proved to have different effects in relation to the best one in terms of outcome, though this did not result always in an increase of the effect for superior stages: this can be derived from the number of events for the 4 types of stage being not perfectly balanced. An increase of the sample size, as well as for the highest level of the WHOPS indicator, could provide further details in this regard.
NOS subtype of NSCLC has proven to be deadlier than the SCC subtype, in comparison to the adenocarcinoma baseline, for all three models.
In conclusion, both flexible multiplicative-additive and additive hazards models provide a useful tool for investigating if time-varying effects are present in the data, especially if the Cox model reveals weaknesses in relation to the proportional hazards assumption. This is valid also when, even after adequately interpolating a covariate's functional form, the model isn't still able to produce reliable estimates.
Starting from a nonparametric framework which assumes all covariate effects to be time-varying, these can later be tested for being time-constant. If, such as in this case, some components appear to have their effect as not constant over time, the model might be simplified step by step leading then to a semi-parametric model which also results

in a simpler interpretation. These evidences also come from the fact that the different approaches implemented throughout this dissertation have consequently resulted in different estimations of the survival probability, as it was shown in Figures 20 and 21.

Code

```
# LIBRERIE
library(survival)
library(survminer)
library(timereg)
library(reshape)
library(reshape2)
library(ggplot2)
library(readxl)
library(MASS)
library(GGally)
library(lmtest)
library(mice)
library(naniar)
library(glmnet)
library(ggfortify)

# UPLOAD
dati_V = read_excel("carvalho-prognostic-biomarkers-NSCLC_IMP_Mea_Med_MLRI.xlsx", 2)
View(dati_V)
dim(dati_V)
table(dati_V$RTProtocol)
str(dati_V)

# RENAMING DELLE VARIABILI CON IL CARATTERE DI SPAZIO
colnames(dati_V)[colnames(dati_V) == "Lymph nodes"] = "Nodes"
colnames(dati_V)[colnames(dati_V) == "RT Protocol"] = "RTProtocol"
colnames(dati_V)[colnames(dati_V) == "Total dose (1st)"] = "TotalDose1st"
colnames(dati_V)[colnames(dati_V) == "Total Dose (2nd)"] = "TotalDose2nd"
colnames(dati_V)[colnames(dati_V) == "IL 6"] = "IL6"
colnames(dati_V)[colnames(dati_V) == "IL 8"] = "IL8"
colnames(dati_V)[colnames(dati_V) == "Cyfra 21-1"] = "Cyfra211"
colnames(dati_V)[colnames(dati_V) == "WHO-PS"] = "WHOPS"
colnames(dati_V)[colnames(dati_V) == "CA-9"] = "CA9"
colnames(dati_V)[colnames(dati_V) == "FEV1s%"] = "FEV1s"
colnames(dati_V)[colnames(dati_V) == "TLR-4"] = "TLR4"
colnames(dati_V)[colnames(dati_V) == "α2M"] = "a2M"
names(dati_V)
View(dati_V)
dim(dati_V)
str(dati_V)

# IMPUTAZIONE DEI VALORI MANCANTI (MEDIANA, MODA E M.I.C.E.)
sum(is.na(dati_V$age))

sum(is.na(dati_V$CRP))

sum(is.na(dati_V$CEA))
```

```r
sum(is.na(dati_V$stage))
find_mode <- function(x) {
  u <- unique(x)
  tab <- tabulate(match(x, u))
  u[tab == max(tab)]
}
moderes = find_mode(dati_V$stage)
ii = which(is.na(dati_V$stage))
dati_V$stage[ii] =  moderes

sum(is.na(dati_V$TLR4))
shapiro.test(dati_V$TLR4)
boxplot(dati_V$TLR4)
medres = median(na.omit(dati_V$TLR4))
ii =which(is.na(dati_V$TLR4))
dati_V$TLR4[ii] =  medres

sum(is.na(dati_V$VEGF))
shapiro.test(dati_V$VEGF)
boxplot(dati_V$VEGF)
median(na.omit(dati_V$VEGF))

sum(is.na(dati_V$Cyfra211))
shapiro.test(as.numeric(dati_V$Cyfra211))
boxplot(dati_V$as.numeric(Cyfra211))
median(na.omit(dati_V$Cyfra211))

sum(is.na(dati_V$OPN))
shapiro.test(dati_V$OPN)
boxplot(dati_V$OPN)
median(na.omit(dati_V$OPN))

sum(is.na(dati_V$CA9))
shapiro.test(dati_V$CA9)
boxplot(dati_V$CA9)
median(na.omit(dati_V$CA9))

sum(is.na(dati_V$GTV))
shapiro.test(dati_V$GTV)
boxplot(dati_V$GTV)
median(na.omit(dati_V$GTV))

sum(is.na(dati_V$a2M))
shapiro.test(dati_V$a2M)
boxplot(dati_V$a2M)
median(na.omit(dati_V$a2M))

sum(is.na(dati_V$histology))
```

```
find_mode(dati_V$histology)

sum(is.na(as.numeric(dati_V$IL8))) # Post numericizzazione per conversione "<" in NA
shapiro.test(as.numeric(dati_V$IL8))
boxplot(as.numeric(dati_V$IL8))
# inserimento valori arbitrari compresi tra 0 e x per "valore < x"

sum(is.na(as.factor(dati_V$WHOPS)))
find_mode(na.omit(as.factor(dati_V$WHOPS)))

sum(is.na(as.numeric(dati_V$IL6))) # Post numericizzazione per conversione "<" in NA
shapiro.test(as.numeric(dati_V$IL6))
boxplot(as.numeric(dati_V$IL6))
# inserimento valori arbitrari compresi tra 0 e x per "valore < x"

sum(is.na(dati_V$TotalDose2nd)) # **: valori NA/dati eccedenti il 43%

mcar_test(dati_V[,-13])

sum(is.na(dati_V$FEV1s))
shapiro.test(dati_V$FEV1s)
boxplot(dati_V$FEV1s)
View(dati_V)

dati_V$Status = as.factor(dati_V$Status)
dati_V$stage = as.factor(dati_V$stage)
dati_V$histology = as.factor(dati_V$histology)
dati_V$Gender = as.factor(dati_V$Gender)
dati_V$WHOPS = as.factor(dati_V$WHOPS)
dati_V$RTProtocol = as.factor(dati_V$RTProtocol)
imp = mice(dati_V[, -c(1,13,22)], method = c(rep("",7), "midastouch",
rep("",14)), print=FALSE, seed=1234)
names(imp)
dim(imp$data)
imp$imp
imp$imp$FEV1s
imputed_values_FEV1s <- mice::complete(imp)$FEV1s
imputed_values_FEV1s[3]
imputed_values_FEV1s[5]
imputed_values_FEV1s[14]
dati_V$FEV1s = imputed_values_FEV1s
sum(is.na(dati_V$FEV1s))

# FATTORIZZAZIONI E NUMERICIZZAZIONI
dati_V$histology = as.factor(dati_V$histology)
dati_V$stage = as.factor(dati_V$stage)
dati_V$Gender = as.factor(dati_V$Gender)
dati_V$RTProtocol = as.factor(dati_V$RTProtocol)
dati_V$WHOPS = as.factor(dati_V$WHOPS)
```

```r
dati_V$IL6 = as.numeric(dati_V$IL6)
dati_V$IL8 = as.numeric(dati_V$IL8)
dati_V$Cyfra211 = as.numeric(dati_V$Cyfra211)
str(dati_V)
View(dati_V)
sum(is.na(dati_V$`IL 8`))

# CONVERSIONE
# (dei valori "alive" e "dead" di dati_V di Status in 0 e 1)
convert_status <- function(dati_V) {
  dati_V$Status <- ifelse(dati_V$Status == "alive", 0, 1)
  return(dati_V)
}
dati_V <- convert_status(dati_V)
View(dati_V)

# CREAZIONE DELLE VARIABILI AGE E NODES SUDDIVISE IN CLASSI
# Age_cat
dati_V$age_cat <- cut(dati_V$age, breaks = c(40, 50, 60, 70, 80, 90), labels = c("40-50",
"50-60", "60-70", "70-80", "80-90"))
dati_V$age_cat <- as.factor(dati_V$age_cat)

# Nodes_cat
bins = c(0, 1, 4)
labels_lymph <- c("0-1", "2-3-4")
dati_V$Nodes_cat <- cut(dati_V$Nodes, bins, labels = labels_lymph)

# ANALISI ESPLORATIVA
# Fattori
# Table per status (dataset di validazione)
table(dati_V$Status)

# Barplot per Sesso
ggplot(dati_V, aes(x = Gender)) + geom_bar(fill = c("steelblue", "pink"))
+ geom_text(stat = 'count', aes(label = ..count..), vjust = 2) +
  labs(title = "Distribuzione per genere", x = "Genere", y = "Frequenza")

# Barplot per histology (dataset di validazione)
ggplot(dati_V, aes(x = histology)) + geom_bar() + geom_text(stat = 'count',
aes(label = ..count..), vjust = 2) +
  labs(title = "Distribuzione per istotipo", x = "Histology", y = "Frequenza")

# Barplot per stage (dataset di validazione)
ggplot(dati_V, aes(x = stage)) + geom_bar() + geom_text(stat = 'count',
aes(label = ..count..), vjust = 2) +
  labs(title = "Distribuzione per stage tumorale", x = "Stage", y = "Frequenza")

# Barplot per RTProtocol (dataset di Validazione)
ggplot(dati_V, aes(x = RTProtocol)) + geom_bar() + geom_text(stat = 'count',
```

```
aes(label = ..count..),vjust = 2) +
labs(title = "Distribuzione per protocollo RT", x = "Protocollo", y = "Frequenza")

# Barplot per classi d'età (dataset di validazione)
table(dati_V$age_cat)
ggplot(dati_V, aes(x = age_cat)) +
  geom_bar(stat = "count", fill = "steelblue") +
  labs(title = "Distribuzione per classi d'età", x = "Età", y = "Frequenza")

# Barplot per Linfonodi (accorpati in due classi: 0-1 e 2-3-4) (dataset di validazione)
# NOTA: D'ora in poi questa variabile verrà considerata solo come suddivisa in classi
table(dati_V$Nodes_cat)
ggplot(dati_V, aes(x = Nodes_cat)) +
  geom_bar(stat = "count") +
  labs(title = "Distribuzione per linfonodi", x = "Linfonodi", y = "Frequenza")

# Variabili quantitative, dataset di validazione (uni- e bi-variate mediante GGpairs)
View(dati_V)
dati_V_quantitative = dati_V[,c(8,9,10,12,13,14,15,16,17,18,19,20,21,22,23,24,25)]
ggpairs(dati_V_quantitative)

# ANALISI NON PARAMETRICA
# Curva di sopravvivenza generale
sopr_V = survfit(Surv(Survival, Status) ~ 1, data = dati_V)
ggsurvplot(sopr_V, data = dati_V, conf.int = TRUE)

# Curva di sopravvivenza per Sesso
sopr_gender_V = survfit(Surv(Survival, Status) ~ Gender, data = dati_V)
ggsurvplot(sopr_gender_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per Terapia (RTProtocol)
sopr_rtprotocol_V = survfit(Surv(Survival, Status) ~ RTProtocol, data = dati_V)
ggsurvplot(sopr_rtprotocol_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per Istologico
sopr_hist_V = survfit(Surv(Survival, Status) ~ histology, data = dati_V)
ggsurvplot(sopr_hist_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per Stage
sopr_stage_V = survfit(Surv(Survival, Status) ~ stage, data = dati_V)
ggsurvplot(sopr_stage_V, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per WHOPS (**)
sopr_whops_V = survfit(Surv(Survival, Status) ~ WHOPS, data = dati_V)
ggsurvplot(sopr_whops_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per Linfonodi
sopr_nodes_V = survfit(Surv(Survival, Status) ~ Nodes, data = dati_V)
ggsurvplot(sopr_nodes_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)
```

```r
# Curva di sopravvivenza per Linfonodi categoriale
sopr_nodes_V = survfit(Surv(Survival, Status) ~ Nodes_cat, data = dati_V)
ggsurvplot(sopr_nodes_V, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

# Curva di sopravvivenza per classi d'eta
sopr_age_V <- survfit(Surv(Survival, Status) ~ age_cat, data = dati_V)
ggsurvplot(sopr_age_V, data = dati_V, conf.int = FALSE, pval = TRUE, pval.method = TRUE)

# CURVE DI SOPRAVVIVENZA DEI BIOMARKER
# opn
dati_V$opn_cat <- cut(dati_V$OPN, breaks = c(0, 75,
150, 225, 300), labels = c("0-75", "75-150", "150-225", "225-300"))
dati_V$opn_cat <- as.factor(dati_V$opn_cat)
sopr_opn_cat = survfit(Surv(Survival, Status) ~ opn_cat, data = dati_V)
ggsurvplot(sopr_opn_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# ca9
dati_V$ca9_cat <- cut(dati_V$CA9, breaks = c(0,
250, 500, 750, 1000), labels = c("0-250", "250-500", "500-750", "750-1000"))
dati_V$ca9_cat <- as.factor(dati_V$ca9_cat)
sopr_ca9_cat = survfit(Surv(Survival, Status) ~ ca9_cat, data = dati_V)
ggsurvplot(sopr_ca9_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE, xlab)

# il6
dati_V$il6_cat <- cut(dati_V$IL6, breaks = c(0, 10,
20, 30, 40), labels = c("0-10", "10-20", "20-30", "30-40"))
dati_V$il6_cat <- as.factor(dati_V$il6_cat)
sopr_il6_cat = survfit(Surv(Survival, Status) ~ il6_cat, data = dati_V)
ggsurvplot(sopr_il6_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# il8
dati_V$il8_cat <- cut(dati_V$IL8, breaks = c(0, 10, 20, 30, 40), labels = c("0-10", "10-20",
"20-30", "30-40"))
dati_V$il8_cat <- as.factor(dati_V$il8_cat)
sopr_il8_cat = survfit(Surv(Survival, Status) ~ il8_cat, data = dati_V)
ggsurvplot(sopr_il8_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# il8
dati_V$crp_cat <- cut(dati_V$CRP, breaks = c(0, 20, 40, 60, 80), labels = c("0-20",
"20-40", "40-60", "60-80"))
dati_V$crp_cat <- as.factor(dati_V$crp_cat)
sopr_crp_cat = survfit(Surv(Survival, Status) ~ crp_cat, data = dati_V)
ggsurvplot(sopr_crp_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# cea
dati_V$cea_cat <- cut(dati_V$CEA, breaks = c(0, 1.5, 3, 4.5, 6), labels = c("0-1.5",
"1.5-3", "3-4.5", "4.5-6"))
dati_V$cea_cat <- as.factor(dati_V$cea_cat)
```

```
sopr_cea_cat = survfit(Surv(Survival, Status) ~ cea_cat, data = dati_V)
ggsurvplot(sopr_cea_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# cyfra
dati_V$cyfra_cat <- cut(dati_V$Cyfra211, breaks = c(0, 1, 2, 3, 4), labels = c("0-1",
"1-2", "2-3", "3-4"))
dati_V$cyfra_cat <- as.factor(dati_V$cyfra_cat)
sopr_cyfra_cat = survfit(Surv(Survival, Status) ~ cyfra_cat, data = dati_V)
ggsurvplot(sopr_cyfra_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# cyfra
dati_V$a2m_cat <- cut(dati_V$a2M, breaks = c(0, 1, 2, 3, 4, 5), labels = c("0-1",
"1-2", "2-3", "3-4", "4-5"))
dati_V$a2m_cat <- as.factor(dati_V$a2m_cat)
sopr_a2m_cat = survfit(Surv(Survival, Status) ~ a2m_cat, data = dati_V)
ggsurvplot(sopr_a2m_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# sil2r
dati_V$sil_cat <- cut(dati_V$sIL2R, breaks = c(0, 3125, 6250, 9375, 12500),
labels = c("0-3125", "3125-6250", "6250-9375", "9375-12500"))
dati_V$sil_cat <- as.factor(dati_V$sil_cat)
sopr_sil_cat = survfit(Surv(Survival, Status) ~ sil_cat, data = dati_V)
ggsurvplot(sopr_sil_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# tlr4
dati_V$tlr_cat <- cut(dati_V$TLR4, breaks = c(0, 5, 10, 15), labels = c("0-5",
"5-10", "10-15"))
dati_V$tlr_cat <- as.factor(dati_V$tlr_cat)
sopr_tlr_cat = survfit(Surv(Survival, Status) ~ tlr_cat, data = dati_V)
ggsurvplot(sopr_tlr_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)

# vegf
dati_V$vegf_cat <- cut(dati_V$VEGF, breaks = c(0, 75, 150, 225, 300), labels = c("0-75",
"75-150", "150-225", "225-300"))
dati_V$vegf_cat <- as.factor(dati_V$vegf_cat)
sopr_vegf_cat = survfit(Surv(Survival, Status) ~ vegf_cat, data = dati_V)
ggsurvplot(sopr_vegf_cat, data = dati_V, conf.int = F, pval = TRUE, pval.method = TRUE)


# CORRELAZIONI
# correlazione tra variabili fisiche e biomarker
str(dati_V)
daticor_V = na.omit(dati_V[,c(9,10,14,15,16,17,18,19,20,21,22,23,24,25)])
str(daticor_V)
cormat_V <- round(cor(daticor_V),2)
melted_cormat_V <- melt(cormat_V)
ggplot(data = melted_cormat_V, aes(x=Var1, y=Var2, fill=value)) + geom_tile()

# MODELLI DI COX (STANDARD)
```

```
# MODELLO COMPRENSIVO DI TUTTE LE VARIABILI
M1step = coxph(Surv(Survival, Status) ~ age + stage + histology + Gender + WHOPS + Nodes
+ RTProtocol + TotalDose1st + GTV + OPN + CA9 + IL6 + IL8 + CRP + CEA + Cyfra211 + a2M + sIL2R
+ TLR4 + VEGF + FEV1s, data = dati_V)
summary(M1step)
cox.zph(M1step, transform = "log")
cox.zph(M1step, transform = "km")
cox.zph(M1step, transform = "identity")
cox.zph(M1step, transform = "rank")
cox.zph(M1step)

M2 = coxph(Surv(Survival, Status) ~ Gender + CEA + histology + WHOPS + Nodes
+ RTProtocol + GTV + OPN + a2M + sIL2R
+ TLR4, data = dati_V)
summary(M2)
cox.zph(M2, transform = "log")
cox.zph(M2, transform = "km")
cox.zph(M2, transform = "identity")
cox.zph(M2, transform = "rank")
table(dati_V$RTProtocol, dati_V$Status)

#simulazione under null hypothesis
fit.cox<-cox.aalen(Surv(dati_V$Survival, dati_V$Status)~prop(dati_V$Gender)+prop(dati_V$CEA)
+prop(dati_V$histology)+prop(dati_V$WHOPS)+prop(dati_V$Nodes)+prop(dati_V$RTProtocol)
+prop(dati_V$GTV)+prop(dati_V$OPN)+prop(dati_V$a2M)+prop(dati_V$sIL2R)+prop(dati_V$TLR4),
weighted.test=0,pbc);
plot(fit.cox,xlab="Time (years)",ylab="Test process",score=T, specific.comps=2)
plot(fit.cox,xlab="Time (years)",ylab="Test process",score=T, specific.comps=9)
plot(fit.cox,xlab="Time (years)",ylab="Test process",score=T, specific.comps=16)

# VALUTAZIONE DELLE FORME FUNZIONALI DEL MODELLO DI BASE
# Res. martingala vs variabile
# residui martingala sul modello finale M2
residui_martingala <- resid(M2, type="martingale")
plot(dati_V$Nodes, residui_martingala, xlab="Nodes", ylab="Residui Martingala")
plot(dati_V$CEA, residui_martingala, xlab="CEA", ylab="Residui Martingala") # **
plot(dati_V$OPN, residui_martingala, xlab="OPN", ylab="Residui Martingala")
plot(dati_V$age, residui_martingala, xlab="age", ylab="Residui Martingala")
plot(dati_V$sIL2R, residui_martingala, xlab="sil2r", ylab="Residui Martingala")
plot(dati_V$GTV, residui_martingala, xlab="GTV", ylab="Residui Martingala") # **
plot(dati_V$Cyfra211, residui_martingala, xlab="cyfra211", ylab="Residui Martingala") #**
plot(dati_V$TLR4, residui_martingala, xlab="TLR4", ylab="Residui Martingala") #**
plot(dati_V$a2M, residui_martingala, xlab="a2M", ylab="Residui Martingala")
plot(dati_V$TotalDose1st, residui_martingala, xlab="dose1st", ylab="Residui Martingala") #**
plot(dati_V$IL6, residui_martingala, xlab="il6", ylab="Residui Martingala")
plot(dati_V$IL8, residui_martingala, xlab="il8", ylab="Residui Martingala")
plot(dati_V$CRP, residui_martingala, xlab="crp", ylab="Residui Martingala") #**
plot(dati_V$VEGF, residui_martingala, xlab="vegf", ylab="Residui Martingala")
ggcoxdiagnostics(M2, type = "martingale", ggtheme = theme_bw())
```

```r
# Residui di Schoenfeld
# Residui schoenfeld CEA (*)
Mcea = coxph(Surv(Survival, Status) ~ CEA, data = dati_V)
ccea = cox.zph(Mcea, transform = "km")
plot(ccea)
ccea = cox.zph(Mcea, transform = "rank")
plot(ccea)
ccea = cox.zph(Mcea, transform = "identity")
plot(ccea)
ccea = cox.zph(Mcea, transform = "log")
plot(ccea)

# Residui schoenfeld OPN
Mopn = coxph(Surv(Survival, Status) ~ OPN, data = dati_V)
copn = cox.zph(Mopn, transform = "km")
plot(copn)
copn = cox.zph(Mopn, transform = "rank")
plot(copn)
copn = cox.zph(Mopn, transform = "identity")
plot(copn)
copn = cox.zph(Mopn, transform = "log")
plot(copn)

# Residui schoenfeld VEGF
Mvegf = coxph(Surv(Survival, Status) ~ VEGF, data = dati_V)
cvegf = cox.zph(Mvegf)
plot(cvegf)

# Residui schoenfeld Nodes (**)
Mnodes = coxph(Surv(Survival, Status) ~ Nodes, data = dati_V)
cnodes = cox.zph(Mnodes, transform = "km")
plot(cnodes)
cnodes = cox.zph(Mnodes, transform = "rank")
plot(cnodes)
cnodes = cox.zph(Mnodes, transform = "identity")
plot(cnodes)
cnodes = cox.zph(Mnodes, transform = "log")
plot(cnodes)

# Residui schoenfeld Cyfra211 (*)
Mcyf = coxph(Surv(Survival, Status) ~ Cyfra211, data = dati_V)
ccyf = cox.zph(Mcyf, transform = "km")
plot(ccyf)
ccyf = cox.zph(Mcyf, transform = "identity")
plot(ccyf)
ccyf = cox.zph(Mcyf, transform = "rank")
plot(ccyf)
ccyf = cox.zph(Mcyf, transform = "log")
```

```
plot(ccyf)

# Residui schoenfeld WHOPS (*)
Mwhops = coxph(Surv(Survival, Status) ~ WHOPS, data = dati_V)
cwhops = cox.zph(Mwhops)
plot(cwhops)

# Residui schoenfeld Gender (*)
Mgen = coxph(Surv(Survival, Status) ~ Gender, data = dati_V)
cgen = cox.zph(Mgen)
plot(cgen)

# Residui schoenfeld TLR4
Mtlr4 = coxph(Surv(Survival, Status) ~ TLR4, data = dati_V)
ctlr4 = cox.zph(Mtlr4)
plot(ctlr4)

# Residui schoenfeld histology (*)
Mhist = coxph(Surv(Survival, Status) ~ histology, data = dati_V)
chist = cox.zph(Mhist)
plot(chist)

# Residui schoenfeld a2M
Ma2m = coxph(Surv(Survival, Status) ~ a2M, data = dati_V)
ca2M = cox.zph(Ma2m)
plot(ca2M)

# Residui schoenfeld RTProtocol (**)
Mrtp = coxph(Surv(Survival, Status) ~ RTProtocol, data = dati_V)
crtp = cox.zph(Mrtp)
plot(crtp)
?cox.zph

# Residui schoenfeld sIL2R
Msil = coxph(Surv(Survival, Status) ~ sIL2R, data = dati_V)
csil = cox.zph(Msil)
plot(csil)

# Residui schoenfeld GTV (*)
Mgtv = coxph(Surv(Survival, Status) ~ GTV, data = dati_V)
cgtv = cox.zph(Mgtv)
plot(cgtv)

# Residui schoenfeld Age
Mage = coxph(Surv(Survival, Status) ~ age, data = dati_V)
cage = cox.zph(Mage)
plot(cage)

# Residui schoenfeld TotalDose1st
```

```r
Mttd = coxph(Surv(Survival, Status) ~ TotalDose1st, data = dati_V)
cttd = cox.zph(Mttd)
plot(cttd)

# Residui schoenfeld IL6
Mil6 = coxph(Surv(Survival, Status) ~ IL6, data = dati_V)
cil6 = cox.zph(Mil6)
plot(cil6)

# Residui schoenfeld IL8
Mil8 = coxph(Surv(Survival, Status) ~ IL8, data = dati_V)
cil8 = cox.zph(Mil8)
plot(cil8)

# Residui schoenfeld CA9
Mca9 = coxph(Surv(Survival, Status) ~ CA9, data = dati_V)
cca9 = cox.zph(Mca9)
plot(cca9)

# Residui schoenfeld CRP
Mcrp = coxph(Surv(Survival, Status) ~ CRP, data = dati_V)
crp = cox.zph(Mcrp)
plot(crp)

# Residui schoenfeld FEV1s
Mfev = coxph(Surv(Survival, Status) ~ FEV1s, data = dati_V)
cfev = cox.zph(Mfev)
plot(cfev)

# curve di sopravvivenza stimate dai due modelli
soprM2 = survfit(M2, data = dati_V)
ggsurvplot(soprM2, data = dati_V, conf.int = TRUE, pval = TRUE, pval.method = TRUE)

sopr_V = survfit(Surv(Survival, Status) ~ 1, data = dati_V)
ggsurvplot(sopr_V, data = dati_V, conf.int = TRUE)

# MODELLO DI COX ESTESO
M2new = coxph(Surv(Survival, Status) ~ Gender + CEA + histology + WHOPS + Nodes
+ RTProtocol + pspline(GTV, df = 0) + OPN + a2M + sIL2R + TLR4, data = dati_V)
summary(M2new)
soprM2new = survfit(M2new, data = dati_V)

termplot(M2new, term=7, se=TRUE, col.term=1, col.se=1)
termplot(M2new, term=11, se=TRUE, col.term=1, col.se=1)
ptemp <- termplot(M2, se=TRUE, plot=FALSE)
attributes(ptemp)

termplot(M2new, term=11, se=TRUE, col.term=1, col.se=1)
ptemp <- termplot(M2, se=TRUE, plot=FALSE)
```

```
attributes(ptemp)
?pspline

# VALUTAZIONE DELLE FORME FUNZIONALI DEL MODELLO A EFFETTI T.DIP.
res = resid(M2tt, "martingale")
ggcoxdiagnostics(M2tt, type = "martingale", ggtheme = theme_bw())

plot(M2new,xlab="Time (years)",ylab="Test process",score=T, specific.comps=10)

mm = timecox(Surv(Survival, Status) ~ Gender + CEA + histology + WHOPS + Nodes + RTProtocol
+ GTV + OPN + a2M, dati_V)
summary(mm)
names(dati_V)

# MODELLI FLESSIBILI: MODELLO A RISCHI ADDITIVI DI AALEN
library(timereg)
fit_aa = aalen(Surv(Survival, Status) ~ GTV + age + stage + histology + Gender
+ Nodes + RTProtocol
+ CRP + CEA + Cyfra211 + a2M + TLR4 + VEGF, data = dati_V)
summary(fit_aa)

# MODELLO DI MCKEAGUE E SASIENI
fit_ms = aalen(Surv(Survival, Status) ~ const(GTV) + const(age)
+ stage + const(histology) + const(Gender) + Nodes + RTProtocol + CRP + CEA + const(Cyfra211)
+ const(a2M) + const(TLR4) + const(VEGF), data = dati_V)
summary(fit_ms)
plot(fit_ms, what = "survival")
plot(fit_aa,xlab="Time (years)",ylab="Test process",score=T, specific.comps=10)
plot(fit_aa,xlab="Time (years)",ylab="Test process",score=T, specific.comps=13)
plot(fit_aa,xlab="Time (years)",ylab="Test process",score=T, specific.comps=14)
plot(fit_aa,xlab="Time (years)",ylab="Test process",score=T, specific.comps=4)
plot(fit_aa,xlab="Time (years)",ylab="Test process",score=T, specific.comps=12)
plot(fit_ms,score=T,xlab="Time (years)",ylab="Test process")

# MODELLO FINALE DI MCKEAGUE E SASIENI
fit_ms2 = aalen(Surv(Survival, Status) ~ const(GTV) + const(age) + const(stage)
+ const(histology) + const(Gender) + Nodes + const(RTProtocol) + const(CRP) + CEA
+ const(Cyfra211) + const(a2M) + const(TLR4) + const(VEGF),
data = dati_V)
summary(fit_ms2)

# MODELLO MOLTIPLICATIVO ADDITIVO DI COX AALEN
fit_ca = cox.aalen(Surv(Survival, Status) ~ prop(WHOPS) + prop(GTV) + prop(age) + prop(stage)
+ prop(histology) + prop(Gender) + Nodes + prop(RTProtocol) + prop(CRP)
+ CEA + prop(Cyfra211) + prop(a2M) + prop(TLR4) +
prop(VEGF), max.time = 8,Nit = 1000, dati_V)
summary(fit_ca)
cox.surv<-list(time=fit_ca$cum[,1],surv=exp(-fit_ca$cum[,2]))
lines(cox.surv$time,cox.surv$surv,type="s",lwd=2,lty=2)
```

```
plot(fit_ca)
plot(fit_ca,score=T,xlab="Time (years)")
fit = aalen(Surv(Survival, Status) ~ const(GTV) + const(age) + const(stage) + const(histology)
+ const(Gender) + Nodes + const(RTProtocol) + const(CRP) + CEA + const(Cyfra211)
+ const(a2M) + const(TLR4) + const(VEGF), data = dati_V,max.time=8, resample.iid=1)
x0<-c(0,0,1); z0<-c(1,0,0);
delta<-matrix(0,length(fit$cum[,1]),181)
for (i in 1:181) {delta[,i]<-x0%*%t(fit$B.iid[[i]])+fit$cum[,1]*sum(z0*fit$gamma.iid[i,]);}
S0<-exp(- x0 %*% t(fit$cum[,-1])- fit$cum[,1]*sum(z0*fit$gamma))
se<-apply(delta^2,1,sum)^.5
plot(fit$cum[,1],S0,type="l",ylim=c(0,1),xlab="Time (years)",ylab="Survival")
fit_ca_s<-cox.aalen(Surv(Survival, Status) ~ prop(WHOPS) + prop(GTV) + prop(age) + prop(stage)
+ prop(histology) + prop(Gender) + Nodes + prop(RTProtocol) + prop(CRP) + CEA
+ prop(Cyfra211) + prop(a2M) + prop(TLR4) + prop(VEGF),
data = dati_V,max.time=8, resample.iid=1)
x0<-c(0,0,1); z0<-c(1,0,0);
delta<-matrix(0,length(fit_ca_s$cum[,1]),181)
for (i in 1:181) {delta[,i]<-x0%*%t(fit_ca_s$B.iid[[i]])+fit_ca_s$cum[,1]
*sum(z0*fit_ca_s$gamma.iid[i,]);}
S0<-exp(- x0 %*% t(fit_ca_s$cum[,-1])- fit_ca_s$cum[,1]*sum(z0*fit_ca_s$gamma))
se_ca_s<-apply(delta^2,1,sum)^.5
surv_ms2 = aalen(Surv(Survival, Status) ~ const(Gender) + const(RTProtocol)
+ const(WHOPS) + const(GTV) + const(stage) + const(histology) + const(Cyfra211)
+ const(a2M) + const(TLR4) + Nodes + CEA, resample.iid = 1, data = dati_V)
summary(surv_ms2)
sur_ms2 = predict.aalen(surv_ms2, dati_V, uniform = F, unif.bands = F)
plot(sur_ms2, col = "blue", ylab = "Survival", xlab = "Time (Years)")
surv_ca = cox.aalen(Surv(Survival, Status) ~ prop(Gender) + prop(RTProtocol)
+ prop(WHOPS) + prop(GTV) + prop(stage) + prop(histology) + prop(Cyfra211) + prop(a2M)
+ prop(TLR4) + Nodes + CEA, resample.iid = 1, data = dati_V)
summary(surv_ca)
sur_ca = predict.cox.aalen(surv_ca, dati_V, uniform = F, unif.bands = F)
plot(sur_ca, col = "red", ylab = "Survival", xlab = "Time (Years)")
?predict.aalen
M2 = coxph(Surv(Survival, Status) ~ Gender + RTProtocol + WHOPS + GTV
+ stage + histology + Cyfra211 + a2M + TLR4 + Nodes + CEA, data = dati_V)
soprM2 = survfit(M2, data = dati_V)
plot(soprM2)
lines(soprM2, conf.int = T, col = "black", lwd = 1)
lines(soprM2, conf.int = F, col = "black", lwd = 2)
lines(soprM2new, col = "orange", lwd = 1, conf.int = T)
lines(soprM2new, col = "orange", lwd = 2, conf.int = F)
```

References

[1] Gaspar LE et al. Amini A, Yeh N. Stereotactic body radiation therapy (sbrt) for lung cancer patients previously treated with conventional radiotherapy: a review. *Radiat Oncol*, 2014.

[2] Lund et al. The role of osteopontin in inflammatory processes. *J Cell Commun Signal, 3(3-4): 311322.*, 2009.

[3] Nguyen et al. Practical strategies for handling breakdown of multiple imputation procedures. *Emerging Themes in Epidemiology volume 18, Article number: 5*, 2021.

[4] Qiang Ma et al. Identification and validation of key genes associated with non-small-cell lung cancer. *J Cell Physiol*, 234(12):22742-22752, 2019.

[5] Soliman et al. Gtv differentially impacts locoregional control of non-small cell lung cancer (nsclc) after different fractionation schedules: subgroup analysis of the prospective randomized chartwel trial. *Radiother Oncol.*, 106(3):299-304, 2013.

[6] Thurneau et al. Martingale based residuals for survival models. *Biometrika*, 77:147:160, 1990.

[7] Finek J Topolcan O Racek J-Minarik M et al. Fiala O, Pesek M. High serum level of c-reactive protein is associated with worse outcome of patients with advanced-stage nsclc treated with erlotinib. *Tumour Biol*, 36:921522, 2015.

[8] Yang W Gai XD Jia T-Lei YM et al. Fu HY, Li C. Foxp3 and tlr4 protein expression are correlated in non-small cell lung cancer: implications for tumor progression and escape. *Acta Histochem*, 115:1517, 2013.

[9] Araujo A Azevedo A Teixeira AL-Catarino R et al. Gomes M, Coelho A. Il-6 polymorphism in non-small cell lung cancer: a prognostic value? *Tumour Biol*, 36:367984, 2015.

[10] Thurneau Grambsch. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515:526, 1994.

[11] Gray. Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics*, 51:1469:1482, 1995.

[12] Sorensen Grunnet. Carcinoembryonic antigen (cea) as tumor marker in lung cancer. *Lung Cancer*, 76(2):138-43, 2011.

[13] Tibshirani Hastie. Generalized additive models. *Chapmand and Hall*, 199o.

[14] Hofman V Ammadi RE Ortholan C-Bonnetaud C et al. Ilie M, Mazure NM. High levels of carbonic anhydrase ix in tumour tissue and plasma are biomarkers of poor prognostic in patients with non-small cell lung cancer. *Br J Cancer*, 102:162735, 2010.

[15] National Cancer Institute. Seer cancer statistics review. *http://seer.cancer.gov/csr/1975*$_2$011/, 1975 : 2011.

[16] Moeschberger Klein. *Survival Analaysis: Techniques for Censored and Truncated Data*. Springer, 2003.

[17] Klein Kleinbaum. *Survival Analaysis: A Self-Learning Text, Third Edition*. Springer, 2012.

[18] Rosinski Klonecki, Kozek. Aalen oo. a model for non-parametric regression analysis of counting processes. *Springer*, Lecture notes in statistics - 2: mathematical statistics and probability theory:1.25, 1980.

[19] Kosinski. Functional form of continuous variable in cox proportional hazards model. *R Help*, Package: survminer, version 0.4.9.

[20] Ying Lin. *Proceedings of the First Seattle Symposium in Biostatistics*. Springer, 1997.

[21] Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association; 1198:1202*, 2012.

[22] Scheike Martinussen. *Dynamic regression models for survival data*. Springer, 2006.

[23] Marzec Marzec. Goodness of fit inference based on stratification in cox's regression model. *Scandinavian Journal of Statistics*, 20(3):227:238, 1993.

[24] Azzoli CG et al Masters GA, Temin S. Systemic therapy for stage iv non-small-cell lung cancer: American society of clinical oncology clinical practice guideline update. *J Clin Oncol*, 2015.

[25] Chyczewska E Naumnik W. The clinical significance of serum soluble interleukin 2 receptor (sil-2r) concentration in lung cancer. *Folia Histochem Cytobiol*, 39:1856, 2001.

[26] Dunst J Dahl O Schild SE-Noack F Rades D, Setter C. Prognostic impact of vegf and vegf receptor 1 (flt1) expression in patients irradiated for stage ii/ iii non-small cell lung cancer (nsclc). *Strahlenther Onkol*, 186:30714, 2010.

[27] Belani C. Ramalingam S. Systemic chemotherapy for advanced non-small cell lung cancer: recent advances and future directions. *Oncologist*, 2008.

[28] Ramsey. Penalized smoothing splines. *R Help*, Package: survival.

[29] Alfaro C Onate C Martin-Algarra S Perez G et al. Sanmamed MF, Carranza-Rua O. Serum interleukin-8 reflects tumor burden and treatment response across malignancies of multiple tissue origins. *Clin Cancer Res*, 20: 5697707, 2014.

[30] Therneau. Spline terms in a cox model. *R CRAN.*, August 13, 2023.

[31] Grambsch Thurneau. *Modeling Survival Data: Extending the Cox Model.* Springer, 2000.

[32] van Baardwijk et al. Mature results of an individualized radiation dose prescription study based on normal tissues constraints in stages i to iii non-small-cell lung cancer. *J Clin Oncol, 28(8):13806.*

[33] van Baardwijk et al. Concurrent chemo-radiation for nsclc to an individualized mld. *Oncology MR, NLM identifier: NCT00572325.*, 2007.

[34] van Elmpt et al. The pet-boost randomised phase ii dose- escalation trial in non-small cell lung cancer. *Radiother Oncol., 104(1):6771.*

[35] Li B Wang Z Sun H-Zhang P et al. Wang J, Yi Y. Cyfra21-1 can predict the sensitivity to chemoradiotherapy of non-small-cell lung carcinoma. *Biomarkers*, 15:594601, 2010.