Bachelor Degree in Applied Statistics
Universitat Autònoma de Barcelona

Final Degree Project

An alternative to the Proportional Hazard Rates model in survival analysis and its application in the assessment of the cyclophosphamide treatment's effect over the survival pattern of Acute Myeloid Leukemia patients†

Joan Pérez Guallar[a] (author) and Llorenç Badiella Busquets[b] (supervisor)

Project context

This work has been carried out during my job as statistical consultant in the Servei d'Estadística Aplicada of the Autonomous University of Barcelona. Thus, it is a real consulting project that has been performed by the service. In fact, this statistical analysis is part of a research project in which several universities from both Spain and France are involved, and it includes data from hospitals in both countries. In fact, this study will be published in an international scientific journal of health sciences, in which Llorenç Badiella and I appear as co-authors of the statistical analysis. On the other hand, this project has been promoted and co-financed by a multinational from the pharmaceutical industry, which commercializes drugs to slow down the development of leukemia. Therefore, given that there are underlying commercial purposes, a confidentiality process has been carried out: the name of the actual substance of the study has been replaced by a similar substance, that is, cyclophosphamide is not the actual substance that has been assessed, but one with similar effects. In this work though the aim is going further than the plain statistical analysis corresponding to the scientific study mentioned before: on one hand, the theoretical background of the statistical procedures used is explored. On the other hand, the methodology used is contextualized with the purpose of justifying its use in similar situations as the case of concern.

Abstract

In the study of time to event data one of the most widely used technique is the Proportional Hazards Model. Nonetheless, this modeling tool is based on several restrictive assumptions which need to be carefully verified before interpretation of parameters estimates. One of them is the assumption of proportional hazard which results directly from the model formula and means that hazard ratio needs to be constant over time. In this work firstly presents the mathematical background concerning the proportional hazard property. Next,it appraises both methodologies for detecting its trustfulness an alternatives when it is rejected. As measures for detecction the $\log(-\log(S(t))$, where $S(t)$ is the survival function, against the logarithm of the time whereas is considered and the Schonfled residuals are analyzed. Additionally, as modelling alternatives both the Aalen Model and the Proportional Hazards Model with time-interaction terms are applied. Afterwards, the theory displayed is illustrated by mean of a real example: the assessment of the Cyclophosphamide treatment's effect over the survival pattern of Acute Myeloid Leukemia patients. In this analysis, the proportional hazard rates assumption for the variable treatment was rejected, refusing thus the time-homogeneous pattern. There were found confirmatory evidences that the treatment with Cyclophosphamide enlarged the the survival of the patients and other events related to the disease's progression ($p < 0.01$) with respect the standard therapy within the first months of the disease. However, this study came across with the fact that the effect of the Cyclophosphamide fades away as time spends, becoming equivalent to the standard therapy from 12 months onwards ($p < 0.01$).

Keywords: Survival Analysis, Cox Model, non-Proportional Hazards, Aalen Model, time-interaction

[a] *BSc in Applied Statistics, Autonomous University of Barcelona, Cerdanyola, Spain.*
[b] *Servei d'Estadística Aplicada, Autonomous University of Barcelona, Cerdanyola, Spain.*

Versió en català

Context del treball

Aquest treball ha estat dut a terme durant la meva feina com a consultor estadístic al Servei d'Estadística Aplicada de la Universitat Autònoma de Barcelona. Així doncs, es tracta d'un projecte de consultoria estadística real que s'ha realitzat des del servei. Aquesta tasca de fet forma part d'un projecte d'investigació en el qual es troben involucrades diverses universitats tant de l'Estat Espanyol com de França, i engloba dades d'hospitals d'ambdós països. A més, aquest estudi serà publicat en una revista internacional de ciencies de la salut, en el qual tant el Llorenç Badiella com jo hi apareixem com a autors de l'anàlisi estadística. D'altra banda, aquest projecte ha estat impulsat i cofinançat per una multinacional de la indústria farmacèutica, la qual té com a objectiu la comercialització d'un nou medicament que realentitzi el desenvolupament de la Leucèmia. Per aquesta raó, donat que hi ha objectius comercials al darrera de l'estudi, s'ha dut a terme un procés de confidencialització del treball: el nom la substància real de l'estudi ha estat substituït per una substància similar, és a dir, la ciclofosfamida no és la substància real que s'ha estudiat, sino que s'ha estudiat una amb efectes similars. en aquest treball tanmateix, es preté anar més enllà de simplement l'àlisi estadística corresponent a l'estudi mencionat. D'una banda, s'estudien de forma rigurosa els fonaments teòrics dels procediments estadístics emprats i per l'altra, es pretén contextualitzar aquesta metodologia amb l'objectiu de justificar el seu ús en situacions similars a la del cas en estudi.

Resum

En l'estudi del temps transcorregut fins a l'esdeveniment una de les tècniques més àmpliament utilitzades és el Model de Riscos Proporcionals. No obstant això, aquesta eina de modelització està basada en diversos supòsits restrictius que necessiten ser verificats acuradament abans de poder interpretar les estimacions dels paràmetres. Un d'ells és l'Assumpció de Riscos Proporcionals que es deriva directament de la fórmula model i essencialment es tradueix en què el quocient de riscos ha de ser constant en el temps. En aquest treball, en primer lloc es presenta la formulació matemàtica referent a la propietat de Riscos Proporcionals. A continuació, es presenten tant metodologies enfocades a la detecció com alternatives quan aquesta no es compleix. Com a eines de detecció considerem la representació gràfica de la funció $log(-\ logS(t))$ (sent $S(t)$ la corba de supervivència) i l'estudi dels residus de Schonfeld. D'altra banda, com a alternatives quan la hipòtesi es rebutjada, presentem el model d'Aalen i el model de riscos proporcionals amb termes d'interacció amb el temps. Posteriorment, la teoria s'il·lustra per mitjà d'un exemple real: l'avaluació de l'efecte del tractament amb ciclofosfamida sobre el patró de supervivència dels pacients amb Leucèmia Mieloide Aguda. En aquesta anàlisi, es rebutja la propietat de riscos proporcionals, descartant així el patró d'homogeneïtat en el temps. D'altra banda, es van trobar evidències significatives que el tractament am cilofosfamida allargava la supervivència dels pacients i retardava altres esdeveniments relacionats amb la progressió de la malaltia, en comparació amb la teràpia estàndard ($p < 0.01$). Tanmateix, també es va trobar que l'efecte de la ciclofosfamida s'esvaeix a mesura que passa el temps, arribant a ser equivalent a la teràpia estàndard a partir dels 12 mesos. ($p < 0.01$).

Paraules clau: anàlisi de la superviència, model de Cox, riscos no proporcionals, model d'Aalen, interaccions amb el temps.

Versión en castellano

Contexto del trabajo

Este trabajo ha sido llevado a cabo durante mi labor como consultor estadístico en el Servei d'Estadística Aplicada de la Universitat Autònoma de Barcelona. Así pues, se trata un proyecto de consultoría real que se ha realizado desde el servicio. Esta tarea, de hecho, forma parte de un proyecto de investigación en el que se encuentran involucradas varias universidades tanto de España como de Francia, y engloba datos de hospitales de ambos países. Además este estudio será publicado en una revista internacional de ciencias de la salud, en el que tanto Llorenç Badiella como yo aparecemos como autores del análisis estadístico. Por otra parte, este proyecto ha sido impulsado y cofinanciado por una multinacional de la industria farmacéutica, la cual tiene como objetivo la comercialización de un nuevo medicamento que realentice el desarrollo de la Leucemia. Por ello, dado que hay objetivos comerciales detrás del estudio, se ha llevado a cabo un proceso de confidencialización del trabajo: el nombre la sustancia real del estudio ha sido sustituido por una sustancia similar, es decir, la ciclofosfamida no es la sustancia real que se ha estudiado, sino que se ha estudiado una con efectos similares . En este trabajo sin embargo se pretende ir un poco más alla del propio análisi estadístico correspondiente al estudio científico mencionado. Por un lado, se estudian de forma rigurosa los fundamentos teóricos de los procedimientos estadísticos empleados y por otro, se pretende contextualizar esta metodología con el fin de justificar su uso en situaciones similares a la del caso en estudio.

Resumen

En el estudio de los datos de tiempo hasta un evento dado, una de las técnicas más utilizadas es el Modelo de Riesgos Proporcionales. Sin embargo, esta herramienta de modelado se basa en varias hipótesis restrictivas que deben ser cuidadosamente verificadas antes de la interpretar los parámetros estimados. Una de ellas es la asunción de riesgo proporcional que resulta directamente de la fórmula del modelo y significa que la razón de riesgo debe ser constante en el tiempo. En este trabajo se presenta en primer lugar el trasfondo matemático relativo a la propiedad de riesgo proporcional. A continuación, se evalúan ambas metodologías para detectar su confiabilidad y alternativas cuando se rechaza. Como herramientas de detección consideramos la gráfica de la función $log(-logS(t))$ (siendo $S(t)$ la curva de supervivencia) contra el logaritmo del tiempo y el análisis de los residuos de Schonfeld. Por otra parte, como alternativas cuando dicha hipótesis es rechazada se consideran el modelo de Aalen y el modelo de riesgos proporcionales con términos de interacción con el tiempo. Posteriormente, la teoría mostrada se ilustra por medio de un ejemplo real: la evaluación del efecto del tratamiento con ciclofosfamida sobre el patrón de supervivencia de pacientes con Leucemia Mieloide Aguda. En este análisis, se rechaza la hipótesis de tasas de riesgo proporcional para la variable tratamiento, descartando así un patrón de homogeneidad en el tiempo. Se encontraron evidencias significativas de que el tratamiento con ciclofosfamida aumentó la supervivencia de los pacientes y retrasó otros eventos relacionados con la progresión de la enfermedad ($p < 0,01$), en comparación con la terapia estándar en los primeros meses de la enfermedad. Sin embargo, también se demostró que el efecto de la ciclofosfamida desaparece conforme el tiempo pasa, convirtiéndose en equivalente a la terapia estándar a partir de los 12 meses ($p < 0,01$).

Palabras clave: análisis de la supervivencia, modelo de Cox, riesgos no proporcionales, modelo de Aalen, interacciones con el tiempo.

Contents

1 Introduction **5**

1.1 The proportional hazard rates assumption in survival analysis 5

1.2 Study case: Assessing the cyclophosphamide treatment's effect over the survival pattern of Acute Myeloid Leukemia patients 5

2 Mathematical background **6**

2.1 The Survival Function 6

2.1.1 The log-rank and Wilcoxon tests 7

2.2 The hazard function and the Proportional Hazards Model 8

2.3 The Proportional Hazard Rates Regression Model 8

2.4 Spreading's analysis of the PH testing praxis 9

2.5 Testing the proportional hazard rates assumption 11

2.5.1 The log-cumulative hazard plot 11

2.5.2 The Schonfeld residuals 12

2.6 Alternatives to the proportional hazard rates model 13

2.6.1 Time-dependent covariates . . 13

2.6.2 The Aalen's Nonparametric, Additive Hazard Model 13

3 Study case: asssessing the Ciclophosphamide treatment's effect over the survival pattern of AMl **14**

3.1 Description of the study 14

3.2 Results 17

3.2.1 Descriptive summary: Demographic and Clinical data . . . 17

3.2.2 Potential risk factors related to Death Event at 1 year 17

3.2.3 Baseline Analysis - Treatment: Demographic and clinical data against Treatment 18

3.2.4 Exploratory Analysis: Bivariate Study 19

3.2.5 Primary Analysis: comparison CYC treatment vs ST 19

3.2.6 Proportional hazard rates hypothesis ascertainment 21

3.2.7 Alternative to proportional hazard rates model 22

3.2.8 Assesing the interaction between treatment and Karyotype 24

3.2.9 Secondary response variables analysis 24

4 Discussions **26**

4.1 The PH assumption in the Cox Model . 26

4.2 Study case 27

5 Appendix: Codes **28**

5.1 R codes 28

5.2 SAS codes 30

1 Introduction

1.1 The proportional hazard rates assumption in survival analysis

In statistics (and in science in general) the most straightforward methodologies for modelling data are always wanted. However, is is probably true to say that the easier is a technique, the more underlying assumptions requires. One such an example is the linear regression model: it is very simple but its validity strongly relies on the assumption of a response variable normally distributed. In this paper we consider a similar situation (more unknown than the previous example though): *the proportional hazard rates assumption in the Proportional Hazard Rates model*. The ignorance of this issue when modelling data might jeopardize the result's validity.

Survival Analysis is the branch of statistics that comprises the set of methodologies devoted to the modelling of time-to-event data. That is to say, an event of interest is fixed and the time from a set origin until the event occurs is measured. Survival Analysis includes a large number of different techniques and spans over a wide range of application levels. No one would dispute the fact that one of the most useful modelling tool in this field, attractive probably by both its easy application and easy interpretation, is the Proportional Hazard Rates Model. *

In a similar way as the validity of linear regression model depends on the normality in the response variable, this technique also depends on an underlying premise: *the proportional hazard rates property*. Most of the studies assume that this hypothesis is fulfilled without carrying out properly testing procedures. If this assumption turns out to be false, and this fact is not taken into account, the results arising from the analysis may lack validity. In case that the proportional hazard rates hypothesis is rejected, other modelling methods must be used. This work aims to answer questions as: What does exactly this property mean? What can be done in case that this property is rejected? Is the level of awareness among the scientific community regarding this issue wide enough? That is to say, it aims to tackle the situation of non-proportional hazard rates in survival analysis from a rigorous standpoint. Broadly, this work is conformed by two distinct parts.

The firsts sections of this work present the mathematical background for approaching the situation accurately. It starts by a brief summary of survival analysis (which obviously includes a presentation of the Cox Regression Model) and a rigorous definition of the proportional hazard rates property. Subsequently, procedures for testing whether the last hypothesis is fulfilled are presented. Next, modelling alternatives for handling data when this premise is rejected are displayed. On top of that, the level of awareness of the proportional hazard rates property testing praxis among the scientific community is assessed by means of the search engine Pubmed (see [18]). This last part pointed out a low level of awareness.

The remainder of the project is devoted to the application of the theoretical methodology depicted in the previous sections to an authentic statistical analysis. This consisted on real study that, as explained in the context of the project, it will be published in a scientific journal. Further information can be found in the next section.

1.2 Study case: Assessing the cyclophosphamide treatment's effect over the survival pattern of Acute Myeloid Leukemia patients

The cancer is one of the main death causes around the world, being the second cause of death right after the cardiovascular diseases. It is estimated that only in the US more than 1.6 million new cases will be diagnosed within the 2017, from which almost 600.000 will die. Moreover, it is forecasted a significant increase in the number of cancer cases. Thus, the prevention, diagnosis and treatment of cancer is of major concern

* This model is also called Cox Regression Model. Both names will be indistinctly used throughout this work.

Among all the types of cancer, one of the most common and severe is leukemia. In this work we deal with the Acute Myeloid Leukemia (AML), which is a cancer of the myeloid line of blood cells. It is characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells. AML is the most common acute leukemia affecting adults, and its incidence increases with age. Although AML is a relatively rare disease, accounting for roughly 1.2% of cancer deaths in the United States cite, its incidence is expected to increase as the population ages.

The symptoms of AML are caused by replacement of normal bone marrow with leukemic cells, which causes a drop in red blood cells, platelets, and normal white blood cells. These symptoms include fatigue, shortness of breath, easy bruising and bleeding, and increased risk of infection. Several risk factors and chromosomal abnormalities have been identified, but the specific cause is not clear. As an acute leukemia, AML progresses rapidly and is typically fatal within weeks or months if left untreated. AML is treated initially with chemotherapy aimed at inducing a remission; people may go on to receive additional chemotherapy or a hematopoietic stem cell transplant. A benefit of treatment with Cyclophosphamide (CYC) in AML patients has been suggested in relatively small studies. Our purpose is to establish the roll of CYC in this situation.

When CYC is used to treat cancer, it works by slowing or stopping the growth of cancer cells in the body. The literature on the study of decelerating leukemia drugs shows a variety of approaches. However, it is generally agreed that a suitable procedure for tackling this situation is studying time-to-event data until a fixed event, which represents in some way the progression of the disease. Patients are split in two groups: one group receives the experimental drug and the other one acts as a control group. Broadly speaking, the next steps consists on studying the survival times until the fixed event and evaluating whether the treated group presented significant higher survival times in comparison with the control group. Notice that the control group is usually treated with a standard therapy rather than remaining untreated. For this procedure, the patient's blood is passed through a special machine that removes white blood cells (including leukemia cells) and returns the rest of the blood to the patient. In this paper in fact, the CYC is compared with the standard therapy (ST). In the case on study the main event considered is the overall survival (OS), that is, the death of the patient. However, the following secondary events are also considered: Progression or Free Survival (PFS, progression of the tumor of death), Tumor Progression (TM) and overall survival after progression (P, death after progression). Furthermore, some other explanatory variables are also taken into account: age, gender, IPSS score, Mielodisplasic Syndrome type and 2008 WHO's disease classification.

2 Mathematical background

In this section, the theoretical background required to develop the statistical analysis of this project is presented. Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen. Survival analysis attempts to answer questions such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

Some of the more important concepts in Survival Analysis are presented next.

2.1 The Survival Function

The survival function is the probability that a patient will survive beyond a specified time. That is, the survival function $S(t)$ is defined as

$$S(t) = P(T \geq t) \tag{1}$$

where T is the random variable that represents the time on study. Note that it can be interpreted as the

proportion of subjects that are not suffered the event within the elapsed time.

Notice that

$$S(t) = 1 - F(t), \tag{2}$$

where $F(t)$ is the distribution function T. From equation (2) the following relation can be derived

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= \frac{\mathrm{d}\,(1-F(t))}{\mathrm{d}t} \\ &= -\frac{\mathrm{d}F(t)}{\mathrm{d}t} \\ &= -f(t), \end{aligned} \tag{3}$$

where $f(t)$ is the density function of T.

If there is censored data, the survival function can be estimated through the **Kaplan-Meier estimator**. Let $t_1 < t_2 < \ldots < t_k$ with $k < n$ be the death times of all the individuals, for each t_i we define

- d_i, the number of deaths in the moment t_i
- n_i, the number of individuals a risk at time t_i.

Then, the Kaplan-Meier estimator of $S(t)$ is

$$\widehat{S}(t) = \prod_{\substack{j \\ t_j < t}} \left(1 - \frac{d_j}{n_j}\right) \tag{4}$$

2.1.1 The log-rank and Wilcoxon tests

In the comparison of two groups of survival data, there are a number of methods that can be used to quantify the extent of between-group differences. Two non-parametric procedures will now be considered, namely the log-rank test and the Wilcoxon test.

2.1.1.1 The log-rank test In order to construct the log-rank test, we begin by considering separately each death time in two groups of survival data. These groups will be labelled as group A and group B. Suppose that there are r distinct death times. $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, and that at time $t_{(j)}$, d_{1j} individuals in group A and d_{2j} individuals in group B die., for $j = 1, 2, \ldots, r$. Unless two or more individuals in a group have the same recorded death time, the values of d_{1j} and d_{2j} will either be 0 or unity. Moreover, suppose that there are n_{1j} individuals at risk in the first group just before time $t_{(j)}$, and that there are n_{2j} et risk in the second group. Consequently, at time $t_{(j)}$, there are $d_j = d_{1j} + d_{2j}$ deaths in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk.

Now consider the null hypothesis that there is no difference in the survival experiences of the individuals in the two groups. One way of assessing the trustfulness of this hypothesis is to consider the extent of the difference between the observed number of individuals in the two groups who die at each of the death times, and the number expected under the null hypothesis. Information about the extent of these differences can then be combined over each of the death times.

Note that we can regard d_{1j} as a random variable which can take any value from 0 to $\min\{d_j, n_j\}$. In fact, d_{1j} is a random variable with hypergeometric distribution, and thus, the probability p that the random variable associates with the number of deaths in the first group takes the value d_j is

$$p = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}}. \tag{5}$$

The mean of the hypergeometric random variable d_{1j} is given by

$$e_{1j} = \frac{n_{1j} d_j}{n_j}, \tag{6}$$

so that e_{1j} is the expected number of individuals who die at time $t_{(j)}$ in group A.

The next step is combining the information described above for each death time to give an overall measure of deviation of the observed values of d_{1j} from their expected values under the null hypothesis. The most straightforward way of doing this is to sum the differences $d_{1j} - e_{1j}$ over the total number of death times r

in the two groups. This yields to the test's statistic

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}). \tag{7}$$

Notice that this this is $\sum_{j=1}^{r} d_{1j} - \sum_{j=1}^{r} e_{1j}$, which is the difference between the total observed and expected number of deaths in Group A. This statistic will have 0 mean , since $E(d_{1j}) = e_{1j}$. Moreover, since the death time are independent. the variance of U_L is simply the sum of variances of the d_{1j}. Now, since d_{1j} has hypergeometric distribution, the variance of d_{1j} is given by

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}, \tag{8}$$

so that the variance of U_L is

$$Var(U_L) = \sum_{j=1}^{r} v_{1j} := V_L. \tag{9}$$

Furthermore, it can be shown that U_L has an approximate normal distribution, when the number of death times is not too small. It follows then that,

$$\frac{U_L}{\sqrt{V_L}} \sim \mathcal{N}(0,1), \tag{10}$$

result (10) allows us to perform the hypothesis test.

2.1.1.2 The Wilcoxon test The Wilcoxon test is also used to test the null hypothesis that there is no difference in the survivor functions for two groups of survival data. The Wilcoxon test is based on the statistic

$$U_W = \sum_{j=1}^{r} n_j (d_{1j} - e_{1j}), \tag{11}$$

where, as in the previous section, d_{1j} is the number of deaths at time $t_{(j)}$ in the first group and e_{1j} is defined in expression (6). The difference between U_W and U_L is that in the Wilcoxon test, each difference $d_{1j} - e_{1j}$ is weighted bu n_j, the total nuber of individuals at risk at time $t_{(j)}$. The effect of this is to give less weight to differences between d_{1j} and e_{1j} for those times when the number of individuals still alive is small, that is, at the longest survival times. This statistic is thus less sensitive that the log-rank-test to deviations of d_{1j} from e_{1j} in the tail of the distribution of survival times. the variance of U_W turns out to be

$$V_W = \sum_{j=1}^{r} n_j^2 v_{1j} \tag{12}$$

where v_{1j} is given by equation (8). Therefore, under the null hypothesis, it holds that

$$\frac{U_W^2}{V_W} \sim \chi_1^2. \tag{13}$$

The Wilcoxon test is hence conducted in the same manner as the log-rank test.

Remark: when the proportional hazards assumption is not fulfilled, the Wilcoxon test is more powerful in comparison with the log-rank test.

2.2 The hazard function and the Proportional Hazards Model

The **hazard function** $\lambda(t)$ is defined as the risk of event at time t and it can be interpreted as the instant rate of events. From a Poisson Process approach it may be interpreted as the rate function of the non-homogeneous Poisson Process that counts the occurrence of the events on study. Thereby, it is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{14}$$

Note that, if T is a continuous random variable, considering the following relations holds

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{15}$$

where $f(t)$ are the $S(t)$ are the density function and survival function of T respectively.

2.3 The Proportional Hazard Rates Regression Model

The *Proportional Hazard Rates Regression Model* (also called Cox Regression Model) models the risk function by means of a baseline hazard function and function of a set of covariates. This model assumes that the risk function of the ith individual with covariates values $X_{i1}, X_{i2}, \ldots, X_{ip}$ may be expressed as

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}) \quad (16)$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are constant coefficients to be estimated from the data and $X_{i1}, X_{i2}, \ldots, X_{ip}$ are the values of the covariates for the inidividual ith. Estimtations are obtained then by Maximum Likelihood.

This model however supposes a strong property, the hazard rates are proportional among levels of the covariates. This condition states that covariates are multiplicatively related to the hazard. In the simplest case of stationary coefficients, for example, a treatment with a drug may, say, halve a subject's hazard at any given time t, while the baseline hazard may vary.

Let us consider two any individuals i and j with vector of covariates X_i an X_j. Calculating the quotient between he risks functions we obtain

$$\frac{\lambda_i(t|X_i)}{\lambda_j(t|X_j)} = \frac{\lambda_0(t)\exp(X_i\beta)}{\lambda_0(t)\exp(X_j\beta)}$$

where β is the coefficient's vector of the model. Therfore

$$\boxed{\frac{\lambda_i(t|X_i)}{\lambda_j(t|X_i)} = \exp((X_i - X_j)\beta)}, \quad (17)$$

, that is to say, the quotient between the hazard functions of any individuals is a constant. This property is called the **property of proportional hazard rates for the individuals i and j**. That is to say, the quotient between the hazard functions of any individuals is a constant. On the contrary, if equation (17) does not holds we will say that the proportional hazard rates property is violated. We will henceforth refer at this property as PH property.

Note that this relation will not hold if and only if the vector of coefficients β is not a constant but a function of the time t. Thus, the underlying condition when this property does not hold is that the covariates act in a time-varying manner over the risk function.

Consider next the case on study: a survival study in which each patient has been allocated to one of two groups, corresponding to a standard treatment and a new treatment. The proportional hazard rates property would mean in this case, that the ratio of the hazard of death at time t in one treatment group relative to the other is independent of survival time.

In the following sections techniques for the testing the relation (17) and alternatives for dealing with data not fulfilling this property are displayed.

2.4 Spreading's analysis of the PH testing praxis

When carriying out a study involving the Cox regression model The evaluation and testing of the PH property becomes a fundamental issue. Otherwise the arising results may lack validity.

This section aims to briefly assess and quantify the spreading's degree of the praxis of contrasting whether this premise is fulfilled, among the health sciences scientific studies. Although other knowledge branches also make use of these survival analysis techniques, it is probably true to say that the main ones are the medicine and health sciences in general. Hence, and following the line of this work, this branch has been chosen for being analysed.

Even though survival analysis techniques have been improved in recent years among the health sciences community, the hypothesis of this work is that the praxis of testing the PH assumption is still far from reciving the importance that deserves. This would suggest that in time-to-event studies things are not always being properly done (from a statistical standpoint).

The procedure for performing this analysis is straightforward: firstly quantifying the number of articles that use the Cox Regression Model, and next, appraising how many of them questioned in some way whether the PH assumption was fulfilled. Thereby, two things are required: a large enough medical scientific articles database and a search engine allowing to filter and making queries. Note that some new concepts regarding the testing the Porportional

Hazards assumption and its posterior correction will be used. Further information will be presented in the next sections.

Both the database and the search engine are provided by the National Center for Biotechnology Information (NCBI) of the US. Specifically, the sub-database containing health science articles is PubMed (see [18] for further information). Moreover, by means of boolean operators, filtering queries in PubMed can be made with ease. Finally, it only rest choosing the right words defining the articles wished to be found.

For example, in order to estimate the number of articles in PubMed which used the Cox Regression Model, the following is introduced in the search engine:

```
(Proportional Hazards Model)
 OR (Cox Model)
 OR (Cox Regression Model)
```

where the boolean operators OR indicate the articles containing at least one of the previous keywords are wished to be returned.

Next, the purpose is to quantify how many of them took into account the PH assumption in some way, thus, the following query is performed

```
((Proportional Hazards Model)
OR (Cox Model)
OR (Cox Regression Model) )
AND
(  (time-interaction) OR
(interaction with time)
OR (time-varying)
 OR (time-dependant) OR
  (Aalen) OR
  (Breslow)
 OR (Wilcoxon) OR (Gehan)
 OR (Schonfeld))  )
```

† That is, keywords related to the PH testing are added to the previous search using the link operator `AND` . In this manner, the previous search results are filtered so that there are only kept those articles that deal with the PH testing in some way.

Additionally, this procedure has been carried out distinguishing between three of the more remarkable medical branches with regards of the use of survival analysis when modelling data. These three are oncology, the HIV study and the transplating surgery techniques. The searches in this case are analogues to the ones before, adding though some keywords related to each field using and `AND` operator (filtering thus the articles by branch). The keywords used, which are supposed to cover the most of the articles in the data base in each case, are

- **Oncology:** oncology, tumour ,tumor , cancer, metastasis, leukemia
- **HIV:** HIV, Human Immunodeficiency Virus, immunodeficiency, AIDS, Acquired Immunodeficiency Syndrome
- **Transplant Surgery:** transplantation, transplant, donor, rejection

The results are gathered in table 1.

Table 1 Total number of articles returned by the search engine for each medicine field and overall, along with those that did considered the PH testing in some way in each case (labeled as *Yes*). Furthermore, the proportion of *Yes* have been calculated. Note that each row arises from a different query, hence, the sum of the fields' results does not yield to the overall results.

	Total	Total Yes	% Yes
Oncology	40,033	952	2.38%
HIV	3,075	135	4.39%
Transplant surgery	6,464	186	2.88%
Overall	88,442	2,455	2.76%

From this table it can be seen that, in all the cases, a few proportion of the articles took into account the testing of the PH property. This supports the hypothesis that too little attention is being paid to this issue.

† The words Gehan and Bresolow refer to the Wilcoxon test. The name of the this contrast varies in literature.

Next, pairwise proportion contrasts are performed with the aim of detecting differences in the proportion of articles that took into account the PH testing among branches. The results can be found in table 2

Table 2 p-values of the pairwise comparisons of proportions of articles considering the PH testing among different branches. The Bonferroni correction for multiple comparisons has been applied.

	Onc.	HIV	Transp. Surgery
HIV	< 0.001	-	-
Transp. Surgery	0.107	< 0.001	-
Overall	< 0.001	< 0.001	1.00

From table 1 some conclusions can be pointed out: firstly, the proportion of articles considering the PH testing in the field of HIV has been found to be significantly greater with respect the rest of the groups. In addition the proportion of the overall case has been found to be significantly greater than the proportion in the case of oncology. The rest of comparisons have turned out to be no significant.

Limitations

Even though some conclusions may be successfully derived from this study, the major drawback of this approach is that it has only been quantified the **proportion of articles that did not take into account the validity of the PH assumption among all the articles** of the field. The truly interesting analysis would consist on quantifying **number the articles in which the PH property was not fulfilled, and this was not took into consideration**. In this way, the problematic of the lack of the PH's testing would be directly assessed.

Nonetheless, this second analysis was impossible to carry out within the framework of this project. One one hand, verifying whether the PH assumption was fulfilled and whether this was considered in the articles, would imply to study deeply each of them individually. Therefore, an enormous amount of work would be needed for obtaining a large enough sample size. On top of that, the downloading and consulting of the most of the articles are subjected to subscription, which obviously costs money. Thus, the lack of subscriptions makes the first step impracticable.

2.5 Testing the proportional hazard rates assumption

As explained in section 2.3, a crucial assumption made when using the Cox regression model is the of PH property. We must therefore consider how the validity of this assumption can be assessed.

2.5.1 The log-cumulative hazard plot

In this section, a straightforward plot that can be used in advance of model fitting is described.

According to the Cox regression model, the hazard of death at any time t for the ith individual is given by expression (16):

$$\begin{aligned}\lambda_i(t|X_i) &= \lambda_0(t)\exp\left(\beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right) \\ &= \lambda_0(t)\exp\left(\beta' X_i\right)\end{aligned}$$

where β is the corresponding vector of coefficients. Integrating both sides of this equation over t gives

$$\int_0^t h(u)\mathrm{d}u = \exp\left(\beta' X_i\right)\int_0^t \lambda_0(u)\mathrm{d}u \qquad (18)$$

Defining the **cumulative hazard function** $H(t)$ as

$$H(t) = \int_0^t h(u)\mathrm{d}u, \qquad (19)$$

equation (18) can be rewritten as

$$H_i(t) = \exp\left(\beta' x_i\right) H_0(t).$$

Taking logarithms at each side of this equation, we get

$$\log H_i(t) = \beta' x_i + \log H_0(t), \qquad (20)$$

from which it follows that differences in the log-cumulative hazard functions do not depend on time. This means that if the log-cumulative hazard functions for individuals with different values of their explanatory variables are plotted against time, the curves so formed will be parallel if the proportional hazards property is accomplished. Otherwise this hypothesis would be rejected. In practice, plotting

the log-cumulative hazard functions against the logarithm of t rather than t itself provides more interpretable plots and so this form plot is commonly used.

However, it turns out that this analysis can be carried out only by means of the survival function. This arises from the fact that the log-cumulative hazard function accomplishes

$$\log H(t) = \log(-\log S(t)) \tag{21}$$

Equation (21) is straightforward to prove:

$$\begin{aligned} -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= -\frac{1}{S(t)}\frac{\mathrm{d}S(t)}{dt} \\ &= \frac{f(t)}{S(t)} \\ &= \lambda(t) \end{aligned}$$

where in the last two equallities expressions (3) and (15) have been used respectively. Next, integrating both sides of relation above we obtain

$$-\log S(t) = \int_0^t \lambda(t) = H(t)$$

from which directly follows equation (21).

To use this plot, the survival data are first grouped according to the levels of one or more factors. If continuous variables are to feature in this analysis, their values will first need to be grouped in some way to give a categorical variable. The Kaplan-Meier estimate of the survival function of the data in each group is then obtained. Next, the curves $\log(-\log S(t))$ have to be plotted against $\log t$. If the proportional hazard rates across the different groups is a likely premise, then this plot will yield parallel curves. Otherwise this hypothesis should be rejected.

2.5.2 The Schonfeld residuals

In this section a method for determining the validity of the PH rates based on model's residuals is explored. These were proposed by Schonfeld (1982).

An important property of these residuals is that there is not a single value of the residual for each individual, but a set of values, one for each explanatory variables included in the Cox regression model.

The ith Schonfeld residual for the jth explanatory variable in the model, is given by

$$r_{Sij} = \delta_i(x_{ij} - \widehat{a}_{ij}) \tag{22}$$

where x_{ji} is the value of the jth explanatory variable, $j = 1, 2, \ldots, p$, for the ith individual in the study and,

$$\widehat{a}_{ij} = \frac{\sum_{l \in R(t_i)} x_{il} \exp\left(\widehat{\beta}' X_l\right)}{\sum_{l \in R(t_i)} \exp\left(\widehat{\beta}' X_l\right)} \tag{23}$$

where $R(t_i)$ is the set of all individuals at risk at time t_i. Note that non-zero values of these residuals only arise for uncensored observations. In addition, if the largest observation in a sample of survival times is uncensored, the value of $\widehat{a}_{ij}$ for that observation, from equation (23), will be equal to x_{ij} and thus $r_{Sij} = 0$.. To distinguish residuals that are genuinely zero from those obtained from censored observations, the latter are usually expressed as missing values.

It turns out that a scaled version of the Schonfeld residuals, proposed by Grambsch and Therneau (1994), is more effective detecting departures from the assumed model. Let the vector of Schonfeld residuals for the ith individual be denoted by $r_{Si} = (r_{S1i}, r_{S2i}, \ldots, r_{Spi})$. The scaled, or wheighted Schonfeld residuals, r^*_{Sji} are then the components of the vector

$$r^*_{Si} = r\mathrm{Var}\left(\widehat{\beta}'\right) r_{Si} \tag{24}$$

where r is the number of deaths among the n individuals, and $\mathrm{Var}\left(\widehat{\beta}'\right)$ is the variance-covariance matrix of the parameter estimates in the fitted Cox regression model. Note that these new residuals are not difficult to compute.

The Schonfeld residuals are particularly useful in evaluating the assumption of proportional hazard rates after fitting a Cox regression model. Grambsch and Therenau fact that the ith Schonfeld residual for the

jth explanatory variable in the model accomplishes

$$\mathrm{E}\left(r^*_{Sji}\right) \approx \beta_j(t_i) - \widehat{\beta}_j, \tag{25}$$

where $\beta_j(t)$ is taken to be a time-varying coefficient of X_j, $\beta_j(t_i)$ is the value of the coefficient at the death time t_i, and $\widehat{\beta}_j$ is the estimated value of β_j in the fitted Cox regression model. The proof of expression (25) can be found in
Consequently, a plot of the values of $r^*_{Sij} + \widehat{\beta}_j$ against the death times should give information about the form of the time-dependent coefficient of X_j, $\beta_j(t)$..

Nevertheless, the graphs obtained in this way are usually quite "noisy" and their interpretation is much helped by superimposing a smoothed curve that s fitted to the scatterplot. There are a number of such smoothers that can be obtained, including *smoothing splines*, but the one that is most commonly used and that will be considered in this work, is the *LOWESS* (locally weighted scatterplot smoothing) smoother , proposed by Cleveland and Loader(1979). Further information regarding this technique can be found in

2.6 Alternatives to the proportional hazard rates model

2.6.1 Time-dependent covariates

In section 2.3 the hazard model for an individual was modeled as a function of fixed-time covariates. These are explanatory variables recorded at the star of the study whose values are fixed throughout the course of the study. In this case in general the main goal was to assess the relationship between the risk groups defined by the covariates to the hazard of relapse or death, controlling for possible confounding variables which might be related to relapse or death.

It is possible, nevertheless to consider covariates whose effect over the reponse do depend on the time t. In fact, time-dependent covariates might also be considered, however, this approach is not within the framework of this project.

Even though variables with time-dependent effects are considered, the Cox model may still be used. The consideration of covariates with time-dependent effects is not only a way for testing whether the proportional hazard rates assumptions is fulfilled, but also provides an alternative approach when it is rejected.

When these kind of variables are used to assess the PH assumption, the Cox Model is extended to contain *product terms* (that is, interaction terms) involving the time-independent variable being assessed and some suitable function of time. That is, if X_i is a constant covariate, a time-dependent covariate arising from X_i, $X_i(t)$ is

$$X_i(t) = g(t) \cdot X_i \tag{26}$$

where t is the time on study. On the other hand, if the PH assumption is being evaluated for X_i, a Cox model might be extended to include the variable $X_i(t)$ in addition to X_i. If this new variable turns out to be significant in the model, the PH hypothesis should be rejected for X_i. Note that from a statistical standpoint, this procedure means considering interactions between the variable t and the rest of the covariates. Including the interaction in the model enables interpretation of the parameters that takes into consideration the fact that the covariate's influence on the hazard level is not constant.

It is important to point out that even though a covariate does not have a inherent time-varying pattern, it does not imply that its effect over the risk function is constant.

When no prior time-varying pattern of a covariate is known, as far as the function type is concerned, some authors suggest using logarithm rather than any other function (Quantin, et al., 1996), the others however underline that there is no theoretical reason to choose logarithm as this approach is seen rather as a technical solution that enables to avoid numerical problems (Allison, 1995).

2.6.2 The Aalen's Nonparametric, Additive Hazard Model

The proportional hazards model, discussed in the previous two chapters, assumes that the effects of the

covariates are to act multiplicatively on an unknown baseline hazard function.

Estimation of the risk coefficients was usually based on the partial likelihood. In the proportional hazards model, these risk coefficients were unknown constants whose value did not change over time. In this section, we present an alternative model based on assuming that the covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients in this model are allowed to be functions of time so that the effect of a covariate may vary over time.

In this case though, we have a set of covariates with time-dependent effects , $X(t) = (X_1(t), X_2(t), \ldots, X_p(t))$. We assume that the hazard rate at time t, for an individual with covariate vector $X(t)$, is a linear combination of the $X_k(t)$'s, that is to say,

$$\lambda\left(t|X(t)\right) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) X_k(t). \tag{27}$$

where the $\beta_i(t)$ are the coefficient functions of the time t to be estimated from the data.

As opposed to the proportional hazards model where likelihood bases estimation techniques are used, estimations of the risk coefficients are based on a least squares technique. The derivation of these estimators is based on the Poisson Process approach to survival analysis.

Let us consider data defined as a 3-tuple with the form (T_j, δ_j, X_j), $j = 1, \ldots, n$, where T_j is the on study time, δ_j the event indicator, and $X_j(t) = (X_{j1}(t), X_{j2}(t), \ldots, X_{jp}(t))$ is a vector of dimension p of, possibly, time-dependent covariates.

For individual j, we shall consider, given $X_j(t)$, the following expression for the risk function

$$\lambda\left(t|X_j(t)\right) = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) X_{jk}(t) \tag{28}$$

where $\beta_k(t)$, $k = 1, \ldots, p$ are unknown parametric functions to be estimated. Direct estimations of the $\beta(t)$ are difficult to found in practise. However, the cumulative risk functions $B_k(t)$, defined as

$$B_k(t) = \int_0^t \beta_k(s) ds, \;\; k = 0, 1, \ldots, p, \tag{29}$$

turn out to be easily estimated. Thus, the estimates of $B_k(t)$ are used to extract conclusions of the functions $\beta_k(t)$: by the fundamental theorem of calculus, crude estimates of $\beta_k(t)$ are given by the slope (derivative) of the estimate of $B_k(t)$.

To find the estimates of $B_k(t)$ a least-squares technique is used. We need to define a $n \times (p+1)$ design matrix, $X(t)$ as follows:

For the ith row of $X(t)$ we set:

- $X_i(t) = (1, X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t))$ if inidividual i is at risk at time t
- $X_i(t) = (0, 0, \ldots, 0)$ otherwise.

Let $I(t)$ be a $n \times 1$ vector with ith element equal to 1 if subject i dies at t and 0 otherwise. The least-squares estimate of the vector $B(t) = (B_0(t), B_1(t), \ldots, B_p(t))^t$ is

$$\widehat{B}(t) = \sum_{T_i \leq t} \left(X^t(T_i) X(T_i)\right)^{-1} X^t(T_i) I(T_i). \tag{30}$$

The variance-covariance matrix of $B(t)$ is

$$\widehat{\mathrm{Var}}\left(\widehat{B}(t)\right) = \sum_{T_i \leq t} \left(X^t(T_i) X(T_i)\right)^{-1} X^t(T_i) I^D(T_i) \cdot \left(\left(X^t(T_i) X(T_i)\right)^{-1}\right)^t. \tag{31}$$

3 Study case: asssessing the Ciclophosphamide treatment's effect over the survival pattern of AMI

3.1 Description of the study

The study is briefly described next.

Study Objective

Primary objective

- To analyze and compare the survival pattern of AML patients treated with cyclophosphamide or standrad therapy.

Secondary objective

- To analyze and compare the survival pattern of AML patients with different demographic features.

Study design

Exploratory study, observational, prospective, multicentric.

Blinding/masking method(s)

Not applicable

Randomization method(s)

Not applicable

Treatments

- Cyclophosphamide (CYC)
- Standard Therapy (ST)

Variables and analysis sets

A total of 235 patients were enrolled in the study. Allo them were analyzed.

Primary Response Variable

The primary response variable is the time overall survival time OS.

Secondary Response Variables

The secondary response variables are the times:

- PFS: progression or free survival.
- TTP: time tumor progression.
- P: overall survival after progression.

Explanatory variables

The primary explanatory variable is:

- Treatment

The secondary explanatory variables are:

1. Sex
2. Age
3. Karyotype
   - -7
   - 7q-
   - 7p-
   - Complex

   See [12] for further details.
4. IPSS: International Prognostic Scoring System. It an score that measures the severity of the patient.
   - Int-2: Medium severity.
   - High: High severity.
5. MDS type
   - Novo: primary MDS, no apparent risk factors can be found.
   - Secondary: occurs because of damage to the DNA from chemotherapy or radiation therapy previously given to treat another medical condition.
6. 2008 WHO disease's Classification
   - AML/RAEB2
   - RARS/RCUD/RCMD/ RAEB-1/others

   for further information regarding this variable check out [14].

Analysis sets

A total of 235 patients were enrolled in the study. All patients were analyzed.

Missing values imputation procedures

No missing values imputation procedures were applied.

Statistical Methods

Statistical Analysis

The statistical analysis was performed using R v3.1.2 and SAS® v 9.4, SAS Institute Inc., Cary, NC, USA.

For all statistical tests a nominal significance level of 5% ($P < 0.05$) was applied. No adjustments for multiple tests were performed.

All data spreadsheets, analysis codes and outputs were electronically stored and archived.

Data Management

All data management, reading and listing was performed using both the R v3.1.2 and the SAS® v.9.4., SAS Institute Inc.

Data validation was performed in order to verify data quality. Missing data, data entry errors or out of range values were checked and potential inconsistencies between variables were detected, reported and corrected after consultation with the study investigators.

Primary Response Variable Analysis

Firstly, the potential relationship between the response variable and the explanatory variables was primarily examined by means of bivariate analyses. The survivals functions for each explanatory variable and each group were estimated using the Kaplan-Meyer estimator with the aim of detecting these potential relationships.

Secondly, a collinearity study was carried out successfully, that is, no significant associations among explanatory variables were found. Nevertheless, for the sake of briefness this part is not showing up in this report.

Next, by means of both a plot of the risk functions and a Schoenfeld residuals analysis, the hypothesis of proportional risks between the patients treated with CYC and ST was rejected.

Once this premise was refused, as alternative to the proportional hazards model, the Aalen's Nonparametric, Additive Hazard Model was adjusted with the aim of obtain useful plots.

Then, in order to assess the CYC treatment's time-varying effects, a proportional hazards model with a time interaction term was considered. Finally, by means of this modeling tool, the excess of risk due to the treatment with CYC with respect the standard therapy was estimated for different points of time and grouping for the variable karyotype.

Secondary Response Variable Analysis

The model containing time-interactions effects found in the primary analysis was applied for modeling the secondary response variables.

3.2 Results

The results of the analysis are displayed, interpreted and explained next.

3.2.1 Descriptive summary: Demographic and Clinical data

Firstly, a table of demographic data was presented. For the continious variables its mean(

Table 3 Demographic data

	[ALL] N=235
Age	67.5 (14.9)
Gender:	
Male	144 (61.3%)
Female	91 (38.7%)
Karyotype:	
-7	55 (23.5%)
7q-	38 (16.2%)
7p-	4 (1.71%)
Complex	137 (58.5%)
IPSS:	
Int-2	119 (50.6%)
High	116 (49.4%)
Treatment:	
ST	120 (51.1%)
CYC	115 (48.9%)
MDS type:	
Novo	135 (57.4%)
Secondary	48 (20.4%)
Unknown	52 (22.1%)
WHO 2008 Classification:	
AML/RAEB2	122 (52.1%)
RARS/RCUD/RCMD/ RAEB-1/others	82 (35.0%)
Unknown	30 (12.8%)

3.2.2 Potential risk factors related to Death Event at 1 year

Next, a table with the risk factors related to the OS event at one year from the start is presented.

	No event N=113	Event N=122	HR	p.overall	N
Age	69.4 [58.7;75.4]	72.0 [64.0;78.6]	1.02 [1.00;1.03]	0.032	234
Gender:				0.371	235
Male	65 (57.5%)	79 (64.8%)	Ref.		
Female	48 (42.5%)	43 (35.2%)	0.84 [0.58;1.22]		
Karyotype:				<0.001	234
-7	32 (28.6%)	23 (18.9%)	Ref.		
7q-	30 (26.8%)	8 (6.56%)	0.44 [0.20;0.98]		
7p-	2 (1.79%)	2 (1.64%)	1.48 [0.35;6.28]		
Complex	48 (42.9%)	89 (73.0%)	1.87 [1.18;2.96]		
IPSS:				<0.001	235
Int-2	69 (61.1%)	50 (41.0%)	Ref.		
High	44 (38.9%)	72 (59.0%)	1.93 [1.34;2.77]		
Treatment:				<0.001	235
ST	46 (40.7%)	74 (60.7%)	Ref.		
CYC	67 (59.3%)	48 (39.3%)	0.51 [0.36;0.74]		
MDS type:				0.647	235
Novo	68 (60.2%)	67 (54.9%)	Ref.		
Secondary	22 (19.5%)	26 (21.3%)	1.07 [0.68;1.68]		
Unknown	23 (20.4%)	29 (23.8%)	1.23 [0.80;1.90]		
WHO 2008 Classification:				0.091	234
AML/RAEB2	54 (48.2%)	68 (55.7%)	Ref.		
RARS/RCUD/RCMD/ RAEB-1/others	46 (41.1%)	36 (29.5%)	0.66 [0.44;1.00]		
Unknown	12 (10.7%)	18 (14.8%)	1.09 [0.65;1.83]		

3.2.3 Baseline Analysis - Treatment: Demographic and clinical data against Treatment

A table of the clinical and demographic data stratifiying by treatment is displayed.

	ST N=120	CYC N=115	OR	p.overall	N
Age	73.8 [62.5;78.9]	69.8 [60.0;74.8]	1.00 [0.98;1.02]	0.051	234
Gender:				0.110	235
Male	80 (66.7%)	64 (55.7%)	Ref.		
Female	40 (33.3%)	51 (44.3%)	1.59 [0.94;2.71]		
Karyotype:				0.804	234
-7	29 (24.2%)	26 (22.8%)	Ref.		
7q-	22 (18.3%)	16 (14.0%)	0.81 [0.35;1.88]		
7p-	2 (1.67%)	2 (1.75%)	1.11 [0.11;11.3]		
Complex	67 (55.8%)	70 (61.4%)	1.16 [0.62;2.19]		
IPSS:				0.134	235
Int-2	67 (55.8%)	52 (45.2%)	Ref.		
High	53 (44.2%)	63 (54.8%)	1.53 [0.91;2.57]		
MDS type:				<0.001	235
Novo	61 (50.8%)	74 (64.3%)	Ref.		
Secondary	7 (5.83%)	41 (35.7%)	. [.;.]		
Unknown	52 (43.3%)	0 (0.00%)	. [.;.]		
WHO 2008 Classification:				<0.001	234
AML/RAEB2	47 (39.2%)	75 (65.8%)	Ref.		
RARS/RCUD/RCMD/ RAEB-1/others	43 (35.8%)	39 (34.2%)	. [.;.]		
Unknown	30 (25.0%)	0 (0.00%)	. [.;.]		

3.2.4 Exploratory Analysis: Bivariate Study

The first step for carrying out a multivariate study is a bivariate exploratory analysis. This procedure aimed to figure out which of the considered explanatory variables had a potential effect over the response variable. Thus, survival plots for each explanatory variable and for each group were represented.

Regarding figure 1, some observations can be made. On one hand, it is highly probable that the sex and the age did not have a significant effect over the response variable inasmuch as the survival curves for the different levels of each variable intersect and the confident intervals overlap.

On the other hand, it stands to reason that the variable Karyotype could have a significant effect over the response, being the type 4 the level with a lowest survival average time.

Finally, there is an strong chance that the variable IPSS had a significant effect over the reponse variable (being the level High the one with lowest survival time), inasmuch as the survival curves did not intersect and for some periods the confindent intervals do not overlap.

3.2.5 Primary Analysis: comparison CYC treatment vs ST

The goal of this section is to asses the effect of the treatment over the variable response, that is, if the treatment have statistically significant effect over the PFS time.

The Kaplan-Meyer estimator of the survival curves was calculated and represented, distinguishing between patients treated or not with CYC. Moreover, 95% level confident intervals were plotted for both curves thus, allowing to contrast differences between CYC-treated patients and the non-treated (that is, treated with the ST).

Next, with the purpose of quantifying the extent of between-groups differences, hypothesis tests whether both curves were equal were carried out. The results are displayed in table 4.

The results in table 4, all together with some knowl-

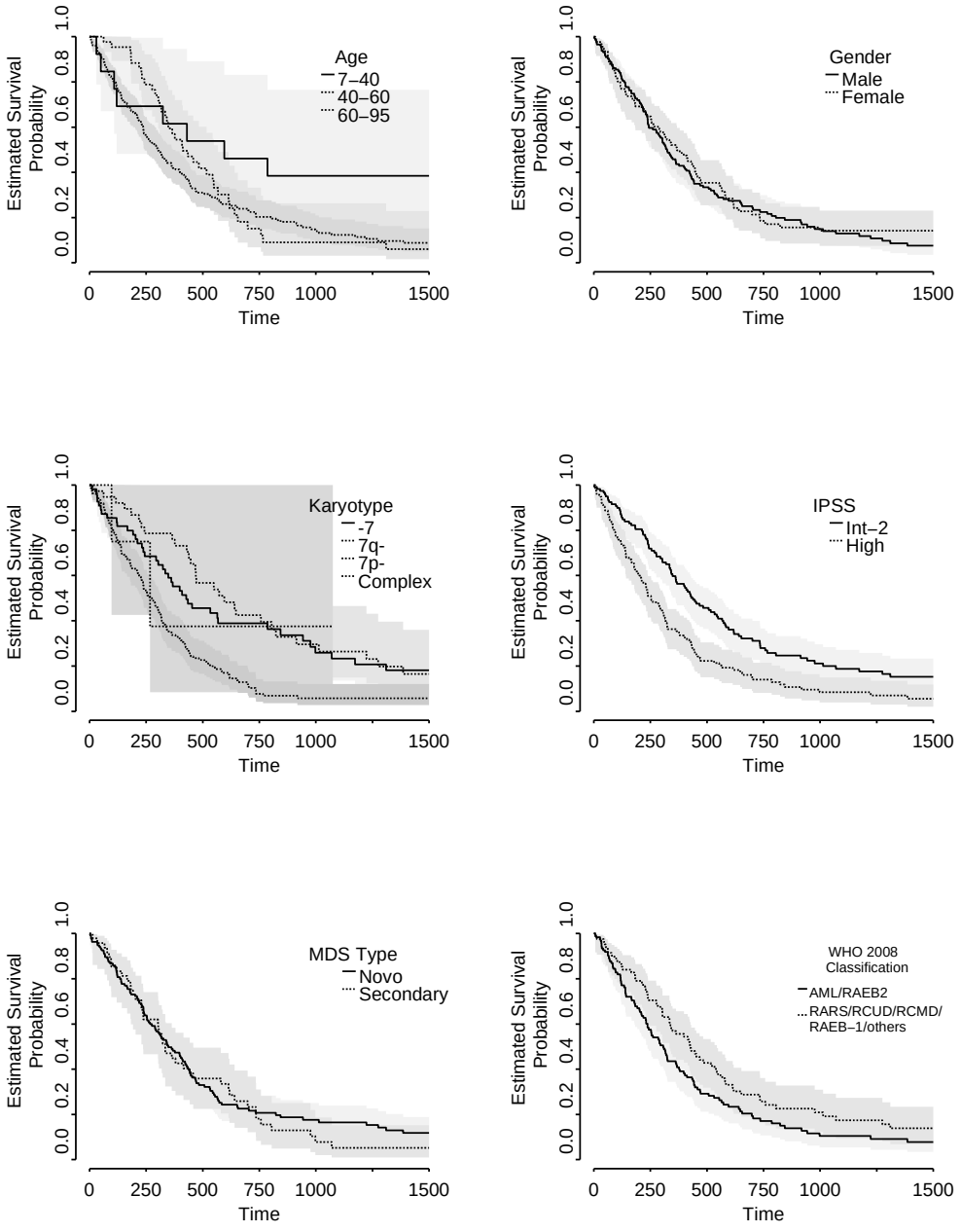

Fig. 1 Survival plots for the study's explanatory variables

Table 4 Results of the tests whether there is equality between the survival curves of each treatment.

test	chisq	df	p.value
Log-Rank	3.4	1	0.063
Wilcoxon	11.5	1	<0.001

edge of each test works, allow to understand the treatment's effect in the response variable.

The log-rank test gives the same importance to all the observations whilst the Wilcoxon test consists in a weighted version of the former. Since the number of observations decreases over time, the Wilcoxon test

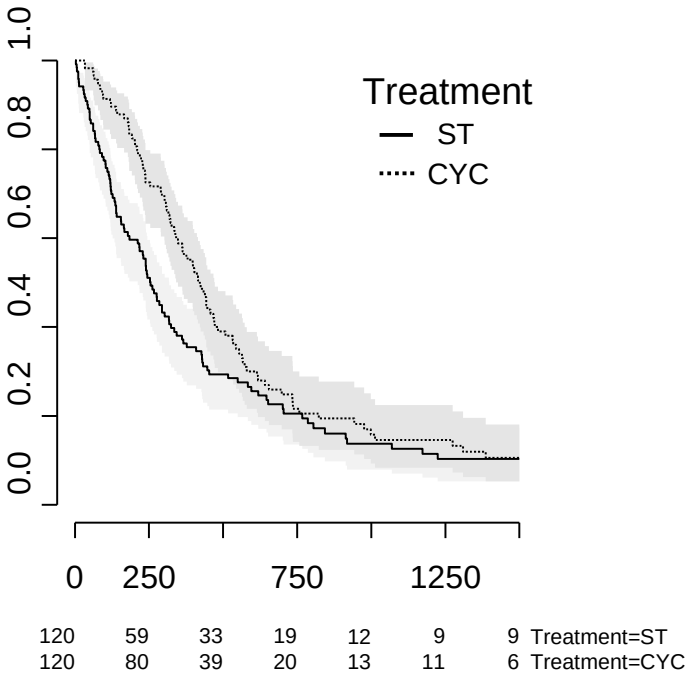

Fig. 2 Estimated survival curves for the treatments STand CYC along with 95% confident intervals.

is more powerful when detecting differences in the early times and the log-rank test is more sensitive to deviations in large times.

Hence, since is has been obtained $p = 0.178$ for the log rank test, it means that there are no globally significant differences in the survival curves. Nonetheless, since it has been obtained a $p < 0.01$ in the Wilcoxon test, it is reasonable to think that there are significant differences between curves for the early stages of the disease.

Thus, it is within the bounds of possibility that the CYC treatment's effect has a time-varying pattern: it seems that it has an increasing effect in the patient's survival time during the early disease's stages but it fades away as time spends.

Moreover, this fact can be corroborated by the figure 2, in which the curve for the patients treated with CYC is below that from those treated with the STand, the survival curves do not intersect and the confidence intervals do not overlap in the early stages whereas they do from approximately 400 days onwards.

3.2.6 Proportional hazard rates hypothesis ascertainment

When modeling the risk function in survival analysis a key assumption is proportional hazards among the levels of the explanatory variables. In the case that concerns us, this property consists in supposing that the risk for those patients treated with CYC is *some times* lower/higher than the risk for those treated with the ST. This is a very important step because of the suitable modeling method for the risk function strongly depends on whether this hypothesis is fulfilled. In fact, the underlying phenomena that might cause a lack of proportionality is a time-varying effect of the covariables. Thereby, contrasting the proportional risk property it will actually be tested the time-homogeneity of the variable's effect.

In this section our goal is to check the truthfulness of this property for the covariable treatment.

Method 1: plotting the curve $\log(-\log(S(t)))$ **vs** $\log t$

First, plotting the curve $\log(-\log(S(t)))$ vs $\log(t)$ for each level of each covariable, a visual analysis was performed. If the proportional hazard rates hypothesis were likely, the obtained lines in each case should be parallel.
Since both lines intersect the proportional hazard rates assumption is not a likely premise.

Method 2: Tests and plots based on the Schoenfeld Residuals

If the model assumptions ‡ two facts must occur

- The residuals must scatter around the 0 and should not show any trend.
- The model's coefficient estimation for the treatment covariable should remain constant over time.

Thereby, the Schonfeld residuals and the parameters estimations were estimated in function of time in order to check those premises.
From figure 4 it could be seen that the variable treatment does not fulfill the criteria above. Hence, it can be concluded that a proportional hazard rates model is not a suitable option when modeling this data.

‡ Constant treatment effect and, thus, proportional hazard rates for this covariable.

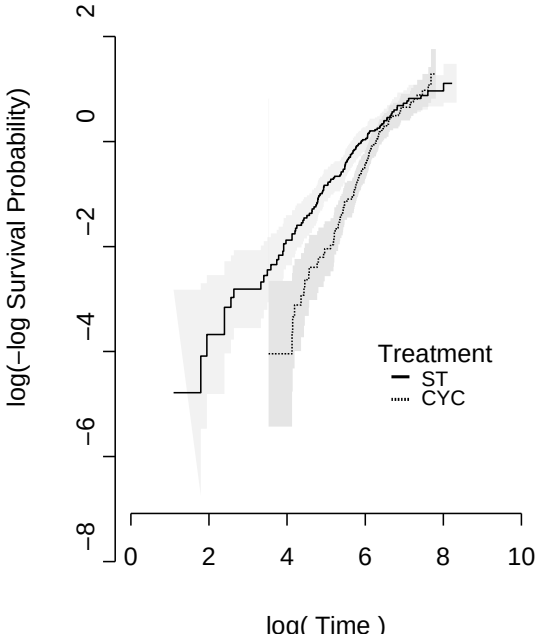

Fig. 3 Plot of the $\log(-\log(\text{S(t)}))$ against $\log t$ for each level of each covariable.

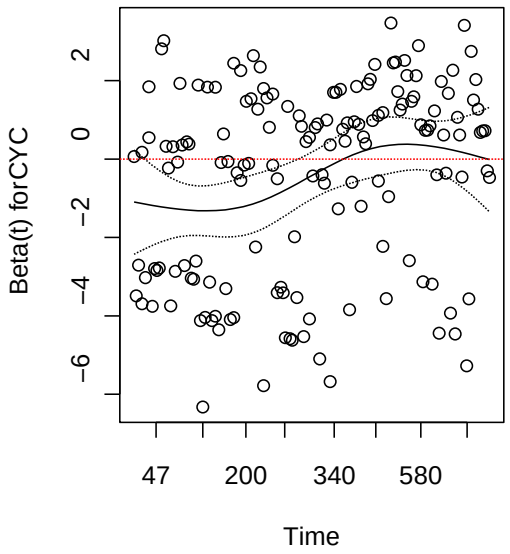

Fig. 4 Sconfeld's residuals plot along with the parameter estimation over time. An horizontal line at zero have been added.

3.2.7 Alternative to proportional hazard rates model

As explained before, f the PH assumption turns out to be false, and this fact is not taken into account, the results arising from the analysis may lack validity. Next, two alternatives are presented.

3.2.7.1 Aalen's Nonparametric, Additive Hazard Model The proportional hazards model, discussed in the previous section, assumes that the effects of the covariates are to act multiplicatively on an unknown baseline hazard function. Moreover, the risks coefficients were unknown constants whose value did not change over time.

Since the required assumptions by the proportional hazards model have been rejected, an alternative model based on assuming that the covariates act in an additive manner on an unknown baseline hazard rate was considered. This model is known as Aalen's nonparametric additive hazard model. In this model the unknown risk coefficients are allowed to be functions of time so that the effect of a covariate may vary over time.

Although this alternative is a nonparametric model, and therefore does not allow to quantify the effect of a covariable over the response variable, the arising plots can be a useful tools when describing the effect of the covariables and its time variation on the response variable.

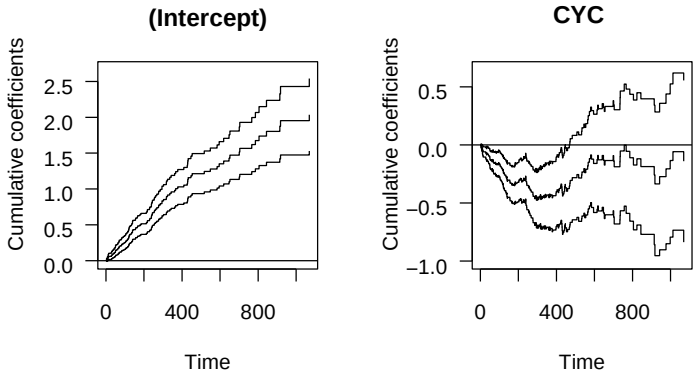

Fig. 5 Cumulative coefficient's estimations for each covariable with 95% confident intervals.

Since of the plots above show cumulative estimates rather than estimates, its interpretation arises from analyzing the increasing/decreasing behavior of the curves instead of the curve itself. It holds that a factor level increases the risk in a given period of time, if for that period the curve is increasing (and vice versa, it decreases the risk if the curve is decreasing).

With this in mind, some qualitative deductions can be

made. On one hand, the cumulative intercept estimate was always increasing thus, according to this model, the basal risk increases over time. On the other hand, the cumulative coefficient estimation corresponding to the patients treated with CYC, showed a decreasing pattern until the day 400 approximately, which turned into a increasing pattern from that day onwards. This fact suggests again a time-varying effect of the CYC treatment over the response variable: first, it decreases the risk whereas it increases it from around 400 days onwards.

It must be pointed out that as the time spends the sample size decreases, whence, the estimations corresponding to early times are more reliable than those from late times. It only takes to have a look at the length of the confidence intervals to realize of this fact. Hence, although the curve corresponding to the CYC treatment starts decreasing again from approximately the 600 days, this measure is not trustworthy, thus, we can't conclude that this treatment decreases the risk again after that moment.

3.2.7.2 Cox regression model with time-interactions. When the proportional hazard rates model assumptions are rejected, an alternative modeling method must be sought. Regarding the previous sections, there are grounds for believing that the cause of this fact is a variation of the CYC's effect over time.

A suitable option that takes into account a time varying effect of the covariables (and thus, a lack of proportionality among hazard rates) turns out to be a Cox Risk Regression Model with an interaction term between these and the time. The results can be found in table 5.
With regards of table 5 some conclusions were pointed out.

Firstly, taking into account the minimmum AIC as model selection criteria, the estimations from simplified model were chosen as the appropriate ones.

Secondly, the CYC treatment effect was found to be significant (p<0.001). Furthermore, the regression coefficient corresponding to the CYC treatment, codified with respect the reference category ST(and thus, showing the excess of risk for patients treated with CYC instead of ST), was found to be $\hat{\beta}_{\text{CYC}} = -3.00$. The negative sign of this figure suggests an overall hazard rate's reduction for those patients treated with CYC with respect those treated with ST.

On top of that, the interaction term between CYC and the time turned out to be statistically significant, corroborating thus, a varying of the CYC treatment's effect over time. This significant interaction brought up the question whether in some of the disease's stages both treatments could become equivalent. In this way, this interaction term was analyzed in detail in order to describe this time-varying effect explicitly. The goal was to quantify the CYC treatment's effect for different stages of the disease.
On account of this fact, the *risk's excess due to the CYC treatment with respect the ST* was estimated for different stages of the disease. This quantity can be understood as a constant that multiplies the risk of an individual (with average values for the rest of the variables) treated with CYC instead of the ST. Therefore, for a given stage of the disease:

Table 5 Multivariate Cox Regression with time-interaction terms. The groups of 2008 WHO Classification RAEB-2/AML and RARS/RCUD/RCMD/RAEB-1/others have been labelled by 1 and 2 respectively.

	First model	Simplified model
Age	0.03 (0.04)	0.012 (0.01)*
Sex - Female	−0.22 (0.18)	
Sex - Male	Ref.	
IPSS-High	2.15 (0.84)*	0.53 (0.15)**
IPSS-Int-2	Ref.	
Karyotype- comp.	0.96 (0.45)	0.57 (0.18)**
Karyotype-7p-	−0.16 (0.35)*	−0.39 (0.73)
Karyotype-7q-	−12.68 (450.00)**	−0.24 (27.85)
Karyotype –7	Ref.	Ref.
MDS Type - Secondary	0.53 (1.15)	
MDS Type - Novo	Ref.	
2008 WHO Clas. - 1	0.69 (1.34)	
2008 WHO Clas. - 2	Ref.	
CYC-Yes	−3.00 (2.34)**	−3.62 (0.91)**
CYC-No	Ref.	Ref.
CYC *Karyotype	−0.17 (0.17)	
CYC* (MDS type)	0.01 (0.52)	
CYC*log(t)	0.2 (0.62)*	0.56 (0.16)**
age* log$t$	0.00 (0.01)	
IPSS*log(t)	−0.28 (0.17)	
Age*log(t)	0.00 (0.00)	
IPSS*CYC*log(t)	−0.04 (0.06)	
Karyotype*CYC	−0.02 (0.12)	
AIC	1227.2.20	1114.677
R^2	0.47	0.52
Max. R^2	1.00	1.00
Num. events	147	201
Num. obs.	172	233
Missings	2	2
PH test	0.00	0.00

$^{**}p < 0.01$, $^{*}p < 0.05$

- If this quantity is near to 0, the CYC treatment causes a huge risk's reduction with respect the ST. Hence, the former is more effective than the latter.
- This risk's reduction decreases as this value is approaching to 1, thus, both treatments start becoming equivalent.
- If this quantity is exactly one this means that the CYC treatment does not make any improvement on risk's reduction with respect the ST. Hence, both treatments are equivalent.
- If this quantity is above 1, the CYC treatment causes a risk's increase with respect the ST. Hence, the former is less effective than the latter.

Regarding the biref explanation above, the interpretation of the figure 6 is as follows:

- At the start, the hazard rate estimate is near to 0. Thus, the treatment with CYC causes a high reduction in the risk with respect de ST.
- This quantity increases within the first and 6 month. Therefore, both treatments start becoming equivalent as time spends.
- Finally, at 1 year, even though the estimate is below one, the differences between the estimate and the unity are not statistically significant. Therefore, all the evidences suggest that once a year have past, the treatment with CYC makes no improvements when reducing the risk (with respect de ST). Hence, both treatments are equivalent.

3.2.8 Assesing the interaction between treatment and Karyotype

Notwithstanding the fact that the interaction term between the Karyotype and the treatment was found to be non-significant when modeling the risk function, this contradicted the investigator's knowledge and experience, whom claimed that the patients with complex Karyotype should respond better to the CYC treatment. Therefore, this interaction was assessed by other means.

First, the observations corresponding to the Karyotype 7p- were removed due to its reduced number. Thereafter, the hazard rates (for those patients treated with CYC with respect those treated with the ST) were estimated and plotted grouping by the variable Karyotype. However, the number of observations corresponding to the types -7 and 7q- were still not large enough, causing thus a huge estimate's standard errors.

An alternative approach consisted on joining the Karyotypes -7 and 7q- and comparing the hazard rate estimates for this joined group with the estimates for the complex karyotype . Thereby, the figure 7 was obtained. Moreover, a table containing the HR's estimates along with 95% confidence intervals and the $p-value$ of the test whether $HR=1$ were presented. Each table corresponds to one karyotype group (CK and non-CK): tables 6 and 7 respectively.

Table 6 HR of the OS event due to the CYC treatment with respect the ST approach for the group **non-complex karyotype**. 95% intervals, and p-values of the test whether the HR is the unit have also been included.

	HR	Lower 95% CI	Upper 95% CI	p-value
Start	0.028	0.001	0.767	0.034
1 month	0.213	0.049	0.935	0.041
3 months	0.412	0.157	1.079	0.071
6 months	0.624	0.300	1.296	0.206
1 year	0.952	0.486	1.866	0.885

Table 7 HR due to the CYC treatment with respect the ST approach for the group **Complex karyotype**. 95% intervals, and p-values of the test whether the HR is the unit have also been included.

	HR	Lower 95% CI	Upper 95% CI	p-value
Start	0.031	0.001	0.711	0.030
1 month	0.184	0.055	0.621	0.006
3 months	0.328	0.166	0.649	0.001
6 months	0.471	0.286	0.778	0.003
1 year	0.682	0.372	1.251	0.217

Logically, the overall pattern showed in the figure 7 is the same that was observed in the figure 6. Nevertheless, although the interaction between karyotype and treatment was found non-significant, it seems that the excess of risk due to the CYC treatment (with respect the ST) within a year for those patients with complex karyotype is subtly smaller than for those with non-complex karyotype. Thus, it stands to reason that the treatment with CYC in the early stages of the disease has a more effective effect in those patients complex karyotype.

3.2.9 Secondary response variables analysis

Even though the mainly goal of this report was to analyze the variable response PFS, that is, the time until the

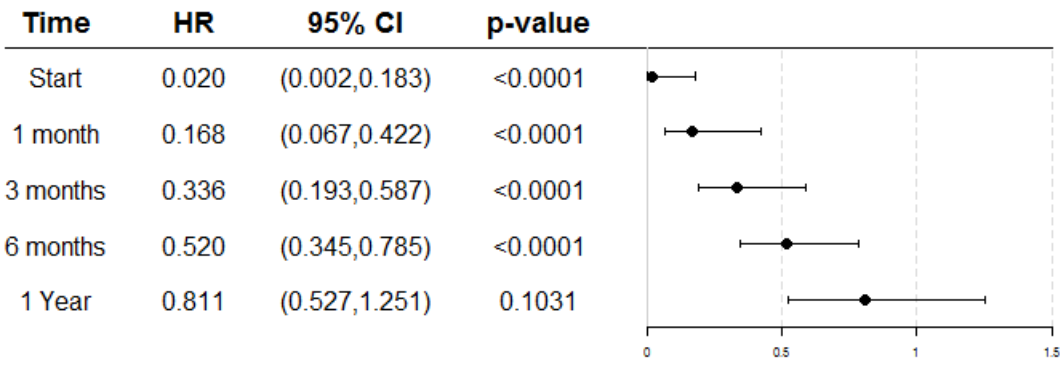

Fig. 6 CYC vs ST HR's evolution over time. The HR estimates are presented along with 95% confidence intervals and the p-value of the test whether HR = 1.

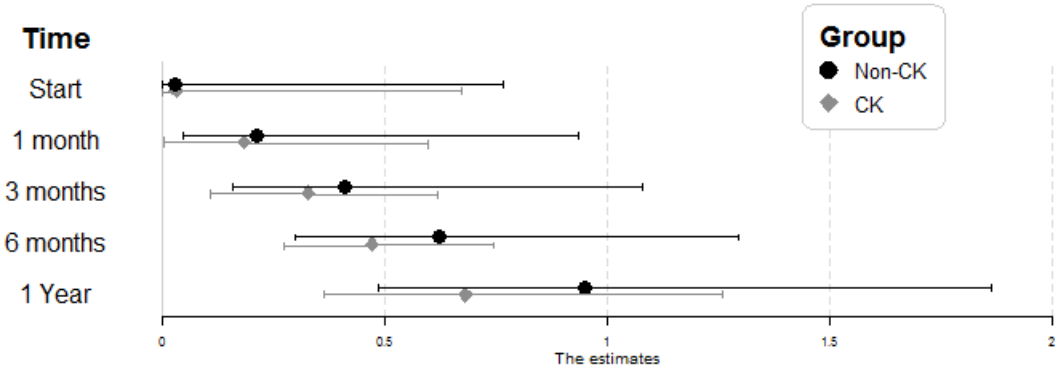

Fig. 7 Excess of risk due to CYC treatment with respect the ST for different disease's stages. The rest of variables have been set at the average values.

event *survival or tumor progression*, some other response variables can provide useful an alternative information. In this section, the effect of the treatment with CYC over the following secondary response variables:

- **PFS** or *Progression or Free Survival* time. It is the time until the event *Progression of the tumor or death of the patient*.
- **TTP** or *time tumor progression*. It is the time until the event *progression of the tumor*.
- **P** or *overall survival after progression*. It is the time between the event *tumor progression* and *death*.

was briefly assessed.

Table 8 Secondary response variables against treatment

	ST N=120	CYC N=115	p.overall	N
PFS	223 [93.8;520]	303 [183;492]	0.003	235
TTP	223 [93.8;520]	303 [183;492]	0.003	235
P	15.0 [1.00;65.0]	78.0 [27.2;149]	<0.001	103

As for the primary response variable, the main goal in this case is modelling the risk function for each of the events associated to the secondary response variables. Similarly to the previous case, the hypothesis of proportional risks among the levels of the factor treatment will be supposed false.

Countinuing with this approach, a suitable procedure was to apply again a Cox Regression model with time-interactions terms. However, for the sake of simplicity and taking into account the results depicted in table 5, the only variables considered in this case were those from the simplified model. Table 9 summarizes de results obtained.

Table 9 Multivariate Cox Regression with time-interaction terms for the events associated to he secondary explanatory varibles.

	PFS	TTP	P
IPSS-High	0.49 (0.15)**	0.63 (0.21)**	−0.02 (0.21)
IPSS- Int 2	Ref.	Ref.	
Kar. -Complex	−0.15 (0.24)	−0.30 (0.38)	−0.51 (0.38)
Kar.- 7p-	−0.26 (0.73)	−13.00 (568.94)	-
Kar. -7q-	0.63 (0.18)**	0.81 (0.26)**	−0.04 (0.26)
Kar. - -7	Ref.	Ref.	
CYC-Yes	−3.77 (0.91)**	−3.69 (0.21)*	−1.58 (0.44)**
CYC - No	Ref.	Ref.	
CYC*log(t)	0.6 (0.16)**	0.66 (0.27)*	0.35 (0.12)**
AIC	1854.03	876.82	736.37
R^2	0.40	0.32	0.26
Max. R^2	1	1	1
Num. events	202	96	97
Num. obs.	234	234	103
Missings	0	1	0

$^{**}p < 0.01$, $^{*}p < 0.05$

The table 9 led to some interesting conclusions. First, it has been found that the treatment with CYC has a significant effect over the risk function corresponding to the three events considered (with $p < 0.01$ in the first and third case and with $p < 0.05$ in the second case).

On the other hand, due to negative sign of all the coeficient's estimates for the treatment CYC, there are confirmatory evidence that the treatment with CYC caused a risk's reduction of the considered events.

Furthermore, as the interaction term between the treatment with CYC was found to be statistically significant in the three cases, this suggested a time-varying effect of the treatment with CYC for the three events considered (with $p < 0.01$ in the first and third case and with $p < 0.05$ in the second case). Additionally, the sign of the estimates corresponding to this term was found to be positives in the three cases, showing thus an increase of the events risk as time spent. That is, in the early disease's stages the treatment with CYC caused a high reduction of the risk (with respect the ST) but this effect faded away as time spent.

To sum up, the pattern followed by the secondary response variables was found to be analogue at that followed by the primary response variable.

4 Discussions

4.1 The PH assumption in the Cox Model

On the one hand, the importance of the proportional hazard rates hypothesis when modelling survival data has been demonstrated. If this property is not fulfilled, the estimates arising should be interpreted carefully. Next, in order to determine the differences between a Cox model with and without time interaction terms (thus, correcting by lack of PH and not correcting) a Proportional Hazards Model was fitted. Such results can be found in table 10.

Table 10 Multivariate Cox Regression

	First model	Simplified model
Age	0.01 (0.01)**	0.01 (0.01)**
Sex - Female	−0.03 (0.15)	
IPSS-High	0.53 (0.15)**	0.53 (0.15)**
IPSS-Int-2	Ref.	
Karyotype- comp.	−0.33 (0.24)	−0.32 (0.24)
Karyotype-7p-	−0.32 (0.73)	−0.30 (0.73)
Karyotype-7q-	−8.68 (337.00)**	−0.17 (22.85)
Karyotype –7	Ref.	0.56 (0.18)**
MDS Type - Secondary	0.63 (1.34)	
MDS Type - Novo	Ref.	
2008 WHO Clas. - 1	0.57 (1.03)	
2008 WHO Clas. - 2	Ref.	
CYC-Yes	−0.46 (0.15)**	−0.47 (0.15)**
CYC-No	Ref.	
AIC	1805.13	1803.17
R^2	0.20	0.20
Max. R^2	1.00	1.00
Num. events	201	201
Num. obs.	233	233
Missings	2	2

$^{**}p < 0.01$, $^{*}p < 0.05$

From tables 5 and 10 it can be seen that, in general, the results arising from both modelling approaches agree. However, note that the Cox Model without interaction terms does not grasp the time-varying pattern of the CYC's effect: the interpretation from this model is merely that the treatment with CYC reduces significantly the risk of

death ($p < 0.001$) with respect the standard approach for any time. Nevertheless, this statement is not true: as explained before, the CYC's effect fades away as time spend, becoming equivalent to the ST approach from the year onwards. Taking into account that the experimental treatments for AML tend to have extremely harmful side effects over the patients, it is highly important to detect when a treatment will have no improvement over the patient health status with respect a more harmless approach. In this case, if the alternatives to the PH model had not been taken into account, the final conclusion would have been that the best procedure for any time would be to give CYC to the patients, causing them therefore, to suffer unnecessary pain and discomfort. In summary, when modelling survival data the PH assumption must be conscientiously tested and corrected if necessary.

With regards of the spreading analysis of the PH testing and correction, it has been found a lack of awareness among the health sciences community. Therefore, there is still much work to do in this context. This project (or one with similar purposes) might be a suitable work to be spread in order to increase the level of attention on this issue.

4.2 Study case

First of all, the differences between treatments in terms of their effect over the survival pattern of the patients were successfully detected by means of the survival curves estimates. Regarding this first analysis, the CYC treatment apparently improved the survival in with respect the ST within the early times.

Secondly, the hypothesis of proportional hazard rates between treatments was assessed and rejected. This fact displayed a lack of homogeneity over time of the treatment's effect.

Afterwards, the Aalen model was applied as an alternative to the proportional hazard rates model. The obtained plot corroborated the time-varying treatment's effect. In addition, it showed a risk's reduction during the early stages of the disease in those patients treated with CYC with respect of the ST.

Next, in order to quantify the effects of all the covariates (and the time's effect as well) a Cox regression model with time-interaction effects was considered. On one hand, the covariates with a significant effect in the risk's variation were found to be the age, the IPSS-High, and the Complex Kariyotype. The three were found to have an increasing effect in the risk's function.

On the other hand, the results showed both a significant effect of the CYC treatment and a significant interaction between time and treatment. Even though the overall CYC effect was a risk's reduction, when the risk's estimates for different periods of time were calculated, the following was detected: in the early stages the treatment with CYC produced a risk's reduction with respect the ST. However, as the time spends, both treatments started becoming equivalent until the point that CYC treatment posed no improvement with respect the ST. This situation is reached approximately in two years.

Then, although the interaction between karyotype and treatment was found to have no effect over the risk's function, taking into account the investigator's experience this interaction was assessed. It turned out that those patients with complex karyotype responded better to the treatment with CYC than those with karyotypes -7 and 7q- within a year [§]. That is,in the early stages of the disease, the risk for those patients with complex karyotype slightly smaller than the risk for those with karyotype -7 and 7q-. Once reached the year, these differences became non-significative.

Finally, after carrying out the primary response variable analysis (the variable PFS), the secondary response variables OS, TTP and P were studied. The obtained results for these variables were analogues to those found in the primary analysis. That is, the treatment with CYC implied a risk's reduction (with respect the ST) of the events associated to the secondary variables, but this effect faded away as time spent.

References

1 Klein, J.P. & . Moeschberger, M. L. *Survival Analysis. Techniques for Censored and Truncated Data*. Springer-Verlag: Statistics for Biology and Health, New York, USA, 2003.

2 Kleinbaum, D. G. *Survival Analysis. A Self-Learning Text* Springer-Verlag, USA. 1995.

3 Badiella, L. & Espinal, A. *Introducción al Análisis de la Supervivencia*. Celgene Publications, Spain. 2010.

4 Hosmer D.& Lemeshow S. *Applied survival analysis. Regression modeling time to event data*. Wiley Series in Probability and Mathematical Statistics. USA. 1999

§ As commented before, since the sample size of the karyotype 7p- was extremely small, no conclusions regarding this group could be made.

5 Fisher, L. D. & Van Belle, G. *Biostatistics. A Methodology for the Health Sciences*. Wiley Series in Probability and Mathematical Statistics.USA. 1993

6 Allison, P.D. *Survival Analysis Using SAS®: A Practical Guide.* . SAS Institute Inc., Cary, NC, USA. 2006.

7 Grambsch, P. M. & Therenau, T. M. *Proportional Hazards Tests and Diagnostics Based on Wheighted Residuals*. Biometrika, Vol 81, No. (Aug 1994), 515-526.

8 Cleveland, W. S. & Loader, C. *Smoothing by Local Regression: Principles and Methods*. AT&T Bell Laboratories, 1993. USA.

9 Wikipedia Contributors. *Leukemia*. Wikipedia, The Free Encyclopedia. Page Version ID: 752612163. 2016.

10 Wikipedia Contributors. *Acute Myeloid Leukemia*. Wikipedia, The Free Encyclopedia. Page Version ID: 751161388. 2016.

11 Wikipedia Contributors. *Cyclophosphamide*. Wikipedia, The Free Encyclopedia. Page Version ID: 751320269. 2016.

12 Janice P, Neoplastic Diseases of the Blood Springer. ISBN 978-1-4614-3764-2. USA 2013co

13 James W. Vardiman, et. al. *The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes.* Blood Journal. 2009 Jul 30;114(5):937-51. doi: 10.1182/blood-2009-03-209262.

14 *NCBI, National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/ . USA

15 *SAS® Suport and Product Documentation*. SAS Institute Inc., Cary, NC, USA.

16 The R project for Statistical Computing http://www.r-project.org/

17 Rstudio: User interface for R http://www.rstudio.com/

18 *PubMed*. US National Library of Medicine . https://www.ncbi.nlm.nih.gov/pubmed/. USA

5 Appendix: Codes

5.1 R codes

Table of demographic data

```
res <- compareGroups( ~ age + sexo +
       cariopronost + IPSSrenew +
       CYC + SMD2p + diagWHO_cat ,
       bd4,include.label <- TRUE,
        max.xlev<-20)
resto <- createTable(res,show.all<-TRUE,
show.n<-FALSE)
caption <- "Demographic data"
```

Potential risk factors related to Progression/Death Event at 6 months

```
bd4$s1 <- pmin(0.5*365,bd4$t1)
bd4$c1 <- 0
bd4$c1[bd4$status1<-<-1
& bd4$s1<-<-bd4$t1 ] <- 1
bd4$s1 <- with(bd4,
Surv(s1, as.integer(c1<-<-1)))
label(bd4$s1)<-"6 months OS Event"

res <- compareGroups( s1 ~ age + sexo +
            cariopronost + IPSSrenew +
 CYC + SMD2p + diagWHO_cat  ,
bd4,include.label <- TRUE, method<-4 )

rest <- createTable(res,show.ratio<-TRUE,
        show.p.ratio<-FALSE,show.n<-TRUE)
caption <- "6 months OS Event"
```

Demographic and clinical data against Treatment

```
res <- compareGroups( CYC ~ age + sexo +
                  cariopronost + IPSSrenew
SMD2p + diagWHO_cat,
bd4,include.label <- TRUE, method<-4)

rest <- createTable(res,show.ratio<-TRUE,
show.p.ratio<-FALSE,show.n<-TRUE)


caption <- "Demographic and clinical data
      against Treatment"
```

Survival curves for the explanatory variables.

¶

```
#the following packages ar required
library(survival)
library(rms)

#Defining the survival time
bd4$T1 <- with(bd4, Surv(t1,
as.integer(status1<-<-1)))
label(bd4$T1)<-"OS Event"

#Plot:
#The age must be classified in categories
```

¶ Since all the plots are constructed analogously, only one case is shown

```
cuts <- c(7,40,60,95)
c1 <- cut(bd4$age,breaks<-cuts)

#Defining an object of class npsruv
surv.age <- npsurv(bd4$T1~c1)

#The plot is constructed through
#the survplot function
survplot(fit  <- surv.age,
    lty<-c(1,2,3),
    lwd<-1,
    conf <- "bands" ,
    xlab <- "Time",
    ylab <- "Estimated Survival \n Probability",
    xlim<-c(0,1500),
    #legend instead of direct label
    label.curves <- FALSE,
    # show only levels, no label
    levels.only  <- F,
    # if label used, abbreviate
    abbrev.label <- F,
    # log(-log Survival) plot
    loglog <- FALSE,
     # log time
    logt   <- FALSE,
     # time increment
    time.inc <- 250,
    # dot grid
    dots  <- F,
    # number at risk
    n.risk  <- F,
    sep.n.risk <- 0.056,
    adj.n.risk <- 1,
    y.n.risk <- -0.24,
    cex.n.risk <- 0.7 )

legend(950,1, bty<-"n",
       title<-"Age",
       c("7-40","40-60","60-95"),
       lty<-c(1,2,3),
       lwd<-1,y.intersp<-0.8,
      x.intersp<-0.2,seg.len<-0.8)
```

Log Rank and Wilcoxon tests

```
 #Log-Rank test
test1<-survdiff(T1 ~ CYC, data<-bd4, rho<-0)
#Wilcoxon test
test2<-survdiff(T1 ~ CYC, data<-bd4, rho<-1)
```

Plot of the $\log(-\log(S(t)))$ against $\log t$ curve

```
surv.CYC <- npsurv(T1 ~ CYC, data <- bd4,
conf.type <- "log-log")
class(p1) <- c(class(p1), "npsurv")

survplot(fit  <- surv.CYC,
         lty<-c(1,4),
         conf <- "bands" ,
         xlab <- "log( Time )",
         xlim<-c(0,10),
         label.curves <- F,
         levels.only  <- FALSE,
         abbrev.label <- FALSE,
         loglog   <- T,
         logt     <- T,
         time.inc <- 250,
         dots     <- FALSE,
         n.risk   <- F,
         y.n.risk <- -0.24,
         cex.n.risk <- 0.7
         )

legend(5,-3.5, bty<-"n",c("BSC","CYC"),cex<-1,
       title<-"Treatment",
       lwd<-2,
       lty<-c(1,4),
       y.intersp<-0.7, x.intersp<-0.3,
       text.width<-4
       ,seg.len<-0.8,xjust<-0)
```

Schonfled residuals

```
cp2 <- coxph(Surv(t1, status1) ~
 age + IPSSrenew +
 cariopronost + CYC + SMD2 +
  diagWHO_cat, data<-bd4)

test <- cox.zph(cp2, transform <- 'rank')
plot(test[6],cex.axis<-1)
abline(h<-0, lty<-3,col<-"red",
xlim<-c(0,5000))
```

Aalen's Nonparametric, Additive Hazard Model

```
#The functions regarding this model
#are  in the timereg package
library(timereg)
aalen.CYC<-aalen(Surv(t1, status1)~CYC,
 bd4,max.time<-1100)
plot(aalen.CYC)
```

Cox Regression Model

```
cp1<-coxph(Surv(t1, status1) ~ age + sexo +
    IPSSrenew + cariopronost +
```

```
    CYC + diag_WHO + SMD,
    data=bd4)

cp2<-coxph(Surv(t1, status1) ~ age + IPSSrenew +
cariopronost + CYC,
data=bd4)

 texreg(list(cp1, cp2),
 custom.coef.names = c("Age",
 "Sex - Female","IPSS-4",
 "Cariotype 2","Cariotype 3",
 "Cariotype 4","Aza"),
float.pos="h",
cap="Multivariate Cox Regression",
caption.above = TRUE,
scriptsize = FALSE,
 custom.model.names=c("First model",
  "Simplified model"),
use.packages=TRUE,
single.row=TRUE,stars = c(0.05, 0.01),
include.adjrs = FALSE)
```

5.2 SAS codes

Cox Model with time-interaction terms

Complete model

```
proc phreg data=a;
class CYC sexo  IPSSrenew cariopronost
   DiagWHO_cat SMD21
   /param=glm DESCENDING;
model t1*status1(0)=   sexo age cariopronost
                       IPSSrenew
                       SMD21
                       DiagWHO_cat  CYC
                       CYCcario CYCSMD21
                       CYCDiagWHO_cat
                       tage tipss tSMD21
                       tDiagWHO_cat tCYC
                       tipsstCYC;
tCYC = CYC*log(t1);
tipss= ipssrenew*log(t1);
tipsstCYC = CYC*ipssrenew*log(t1);
tage=age*log(t1);
CYCcario=cariopronost*CYC;
tSMD21 = SMD21*log(t1);
tDiagWHO_cat = DiagWHO_cat*log(t1);
CYCSMD21 = SMD21*CYC;
CYCDiagWHO_cat =DiagWHO_cat*CYC;
run;
```

Simplified models:

```
*Simplified 1;
proc phreg data=a;
class CYC sexo  IPSSrenew cariopronost
DiagWHO_cat SMD21
 /param=glm DESCENDING;
model t1*status1(0)=   sexo age
                       cariopronost IPSSrenew
                       age cariopronost
                       IPSSrenew SMD21
                       DiagWHO_cat  CYC  tCYC;
tCYC = CYC*log(t1);
run;

*Simplified 2;
proc phreg data=a;
class CYC sexo  IPSSrenew cariopronost
  DiagWHO_cat SMD21
  /param=glm DESCENDING;
model t1*status1(0)=   age cariopronost
                       IPSSrenew  CYC  tCYC;
tCYC = CYC*log(t1);
run;
```

Estimation of risk exces in diferents stages of the disease

```
proc phreg data=a;
class CYC  IPSSrenew cariopronost
                /param=glm DESCENDING;
model t1*status1(0)=   age cariopronost
                       IPSSrenew  CYC  tCYC;
tCYC = CYC*log(t1);
estimate "hr 0d" CYC 1 -1 tCYC
0  0/exp;
estimate "hr 1m" CYC 1 -1 tCYC
 3.4 -3.4/exp;
estimate "hr 3m" CYC 1 -1 tCYC
 4.5 -4.5/exp;
estimate "hr 6m" CYC 1 -1 tCYC
5.193 -5.193/exp;
estimate "hr 12m" CYC 1 -1
tCYC 5.9 -5.9/exp;
run;
```

Assessing the interaction between treatment and karyotype

First, the osbervations of the type 3 are removed. Next the types 1 and 2 are joined

```
*Removing the type 3;
data b;
set a;
```

```
if cariopronost=3 then delete;
run;

*Joining types 1 and 2;
data b;
set b;
if cariopronost=1 | cariopronost=2
then cariopronost=5;
run;
```

Finally, the estimates of the exces of risk due to the CYC treatment for each karyotype are calculated

```
*Type 4;
proc phreg data=b;
where cariopronost=4;
class CYC  IPSSrenew cariopronost
        / param=glm DESCENDING;
model t1*status1(0)=   age cariopronost
                     IPSSrenew  CYC  tCYC;
tCYC = CYC*log(t1);
estimate "hr 0d" CYC 1 -1 tCYC 0  0 / exp;
estimate "hr 1m" CYC 1 -1 tCYC
3.4 -3.4 / exp;
estimate "hr 3m" CYC 1 -1 tCYC
 4.5 -4.5 / exp;
estimate "hr 6m" CYC 1 -1 tCYC
5.193 -5.193 / exp;
estimate "hr 12m" CYC 1 -1 tCYC
5.9 -5.9 / exp;
run;

*Types 1 and 2;
proc phreg data=b;
where cariopronost=5;
class CYC  IPSSrenew cariopronost
/ param=glm DESCENDING;
model t1*status1(0)=   age cariopronost
                       IPSSrenew  CYC  tCYC;
tCYC = CYC*log(t1);
estimate "hr 0d" CYC 1 -1 tCYC 0  0 / exp;
estimate "hr 1m" CYC 1 -1 tCYC
3.4 -3.4 / exp;
estimate "hr 3m" CYC 1 -1 tCYC
4.5 -4.5 / exp;
estimate "hr 6m" CYC 1 -1 tCYC
5.193 -5.193 / exp;
estimate "hr 12m"
```

Secondary response variables analysis

```
*PFS;
proc phreg data=a;
class CYC  IPSSrenew cariopronost
/ param=glm DESCENDING;
model t2*status2(0)=   age cariopronost
 IPSSrenew                        CYC  tCYC;
tCYC = CYC*log(t2);
run;

*Tumor progression;
proc phreg data=a;
class CYC  IPSSrenew cariopronost
/ param=glm DESCENDING;
model t3*status3(0)=   age cariopronost
IPSSrenew  CYC  tCYC;
tCYC = CYC*log(t3);
run;

*Death after progression;
proc phreg data=a;
class CYC  IPSSrenew cariopronost
/ param=glm DESCENDING;
model t4*status4(0)=    age cariopronost
 IPSSrenew CYC  tCYC;
tCYC = CYC*log(t4);
run;
```