

STATISTICAL GUIDELINES FOR EBMT

By M. Labopin and S. Iacobelli

This document is intended to give to EBMT investigators basic methodological recommendations in order to unify the statistical analyses performed in the context of the European Group for Blood and Marrow Transplantation.

The document has been supervised by the statisticians of the EBMT. In addition to them, the authors would like to thank Dr. Richard Szydlo (Imperial College, London, UK) for his useful suggestions and contributions.

EBMT statisticians

Ronald BRAND	Dept. of Medical Statistics Leiden University Medical Center University of Leiden P.O. Box 9604 NL-2300 RC Leiden THE NETHERLANDS	Tel: 31 71 5276734 Fax: 31 71 5276799 e-mail: R.Brand@lumc.nl
Simona IACOBELLI	Dept. of Medical Statistics Leiden University Medical Center P.O. Box 9604 NL-2300 RC Leiden THE NETHERLANDS	Tel: 31 71 5276829 Fax: 31 71 5276799 e-mail: S.Iacobelli@lumc.nl
Myriam LABOPIN	EBMT Central Data Office Faculté de Médecine Hôpital Saint Antoine 27, rue de Chaligny 75571 Paris Cedex 12 France	Tel: 33 1 40469230 Fax: 33 1 40510594 e-mail: labopin@chusa.jussieu.fr
Carmen RUIZ de ELVIRA	EBMT Central Registry Office Department of Hematology Macdonald Buchanan Middlesex Hospital London W1N 8AA UNITED KINGDOM	Tel: 44 20 73809772 Fax: 44 20 723809597 e-mail: C.Ruiz@UCL.AC.UK
Goli TAGHIPOUR	EBMT Central Registry Office Department of Hematology Macdonald Buchanan Middlesex Hospital London W1N 8AA UNITED KINGDOM	Tel: 44 171 3809772 Fax: 44 171 23809597 e-mail: g.taghipour@ucl.ac.uk

Contents

A	General indications for submitting a new study	3
B	Statistical analysis.....	3
1.	Population definition.....	3
2.	Definition of outcomes	3
3.	Variable selection, checks and coding.....	8
4.	Sample description.....	9
5.	Outcomes description: univariate statistical analysis	10
6.	Outcomes analysis: multivariate regression	12
6.1	Introduction.....	12
6.2	Initial variable selection.....	13
6.3	Model identification.....	13
6.4	Model validation	14
6.5	Multivariate survival analysis: choice of the model	14
6.6	Additional features and recommendations for the Cox model	15
7.	Methods for competing risks	16
8.	Matched studies	17
9.	Interpretation of the results	18
C	Presentation of the results	18
1.	Kaplan-Meier and Cumulative Incidence curves.....	18
2.	Cox models	20
D	Further readings	20

A General indications for submitting a new study

A protocol has to be written and submitted to the relevant Working Party, and should include:

- Study contacts:
 - Coordinator of the study
 - Data manager and statistician
- Objectives/Hypotheses
- Background/Rationale/Justification
- Organization of data collection
- Planned statistical analysis
- Schedule
- References

For further indications, see “Guidelines for retrospective studies using the EBMT registries” by P. Ljungman and C. Ruiz de Elvira, available on the EBMT web site (<http://www.ebmt.org/1WhatisEBMT/whatisebmt2.html>).

B Statistical analysis

Once the problem to be investigated is defined, the statistical analysis is usually developed according to the following steps :

- Population definition
- Outcomes definition
- Variables selection and coding
- Sample description
- Outcomes description: univariate statistical analysis
- Outcomes analysis: multivariate regression

1. Population definition

The selection of the sample determines the population to which the conclusions of the study will be generalized (kind of disease, patient characteristics, calendar period and so on).

The population must be selected on the basis of characteristics known at the beginning of the time interval chosen for the outcomes. In other words, if we are studying survival after bone marrow transplantation, we should not use variables which describe the course of the disease after transplantation to define the population under study.

2. Definition of outcomes

Censored variables

In EBMT studies, the main outcomes of interest are events occurring after transplant such as death or relapse. Each event of interest may occur at a variable time post transplant, so in statistical terms it has two components – whether it occurs at all, and if it does, the length of time from transplant to the event. However, in many studies the event of interest is seldom observed in all of the patients. Thus, a patient who has

not yet had the event of interest at the time of analysis, or who is lost to follow-up, is 'censored' at the time of last contact. The mechanism which deals with this type of incompleteness in the data is called "censoring".

The issue of censoring involves actually many statistical problems. As a general rule, to avoid serious statistical bias it is important that the censoring is independent of the future development of the disease or, in other terms, that it is not informative on the prognosis of the patient. An example of such a "non-informative" censoring is the one due to end of follow-up and migration. Nonetheless, some statistical models assume as censoring the occurrence of an event that changes the prognosis (for example, in a Cox model for time to Relapse any non-relapse death is considered as an additional censoring event). Some of these situations will be illustrated below and in section 7.

Outcomes of interest in EBMT

The analysis of censored times to event is usually called "survival analysis", though the variables involved are not only survival times. In this section, we will assess criteria for the definition of outcomes in EBMT studies (endpoints, durations and censoring events), mentioning also the methodologies to be applied, which will be illustrated in sections 4 to 7.

Note, in any case, that it is possible to have slightly different definitions depending on the disease being considered, or that, due to unavailability of information, some statistical method is in fact not applicable. It is therefore recommended the collaboration of the responsible physician with the statistician and/or the data manager of the relative working party in planning the statistical analysis.

Primary outcomes

Overall Survival (OS)

This is the simplest outcome, defined as the probability of survival irrespective of disease state at any point in time. Patients alive at their last follow-up are censored. It is analyzed by the Kaplan-Meier method, Log-Rank Test and parametric or semi-parametric survival models.

Disease-Free Survival (DFS)

DFS is defined as the probability of being alive free of disease at any point in time. Thus, death or disease relapse are treated as events¹. Patients alive and free of disease at their last follow-up are censored. The statistical methods for the analysis of DFS are the same as for OS (Kaplan-Meier curve, Log-Rank Test and survival models).

Relapse Incidence (RI)

RI is defined as the probability of having had a relapse before time t . Death without experiencing a relapse is a competing event. The correct method of analysis is therefore the estimation of the Cumulative Incidence curve, comparable by the Gray Test and, for the multivariate analysis, the application of the proportional hazard

¹ For the analysis of DFS (as well as RI), care must be taken if patients can be transplanted while not in complete remission (CR): the duration of DFS (RI) should be calculated from the date of achievement of CR; unfortunately these data are often missing.

model for the sub-distribution of competing risks, by Fine and Gray (methods suitable for competing risks will be briefly illustrated in section 7).

In studying relapse, sometimes the interest is not only in the estimation of the cumulative incidence curve, but also in the estimation of the hazard ratios for comparing groups of patients. It is therefore common to apply also a survival (Cox or parametric) model considering relapse as an event and death without relapse as a censoring (the response time is given by the minimum between time to relapse and time to death without relapse; as usual, a patient who is alive and free of relapse is also censored).

Non-Relapse Mortality (NRM)

It is defined as the probability of dying without previous occurrence of a relapse, which is a competing event. The same indications as for the analysis of RI apply.

Progression-Free Survival (PFS)

It is defined as the probability of being alive with no indication of disease progression (relapse is considered as progression for patients in CR). It is analyzed by Kaplan-Meier curve, Log-Rank Test and parametric or non-parametric survival models.

Other issues could be considered in a study. An important example is the so-called **Current Disease-Free Survival**, that is the probability of being alive and disease-free at any point in time, including also the situation of being in a subsequent remission after treatment for relapse. This probability is not readily estimable using standard statistical software², and a series of choices for a proper modeling, in the framework of multi-state models (section 6), has to be done.

Another typical example in EBMT studies requiring a choice in accordance with the aim of the study arises in presence of multiple transplantations: the “last follow-up” date and the status can be defined in two different ways:

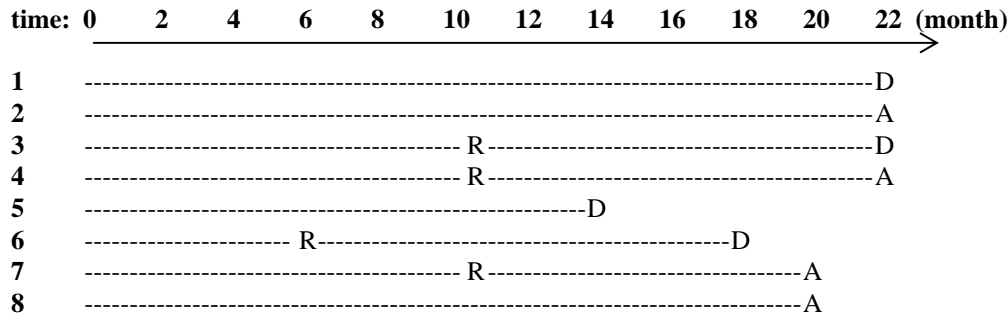
- the date of last contact for the patient, with the last survival status (the unit studied is the patient)
- the date of second transplant, with status defined as “alive at second transplant” (the unit is the graft)

In this as in other similar cases, the choices made have to be indicated both in the proposal of the study and, later, in the report of the results of the statistical analysis.

Example

The following schema represents the course of the disease for 8 patients (R=relapsed, D=dead, A=alive):

² A SAS macro is available for the estimation of Current Disease-Free Survival.



The data are:

Patient number	Relapse (yes/no)	Time to relapse (mo)	Vital status =Status indicator for OS Event=dead=1 Censored=alive=0	Time to death or last contact =Time to event for OS	Time to event for DFS or RI	Status for DFS Event = 1 Censored=0	Status for RI Event = 1 Censored=0 Competing event=2
1	no	-	Dead	22	22	1	2
2	no	-	Alive	22	22	0	0
3	yes	10	Dead	22	10	1	1
4	yes	10	Alive	22	10	1	1
5	no	-	Dead	14	14	1	2
6	yes	6	Dead	18	6	1	1
7	yes	10	Alive	20	10	1	1
8	no	-	Alive	20	20	0	0

Secondary outcomes

Haematopoietic Recovery

The day of engraftment is defined as the first day of 3 consecutive days with a persistent blood cell count above a predefined level:

WBC	$1 \cdot 10^9/l$
PMN	$0.5 \cdot 10^9/l$
Platelets	$50 \cdot 10^9/l$ or $20 \cdot 10^9/l$

Death without recovery is a competing event, while no engraftment at the last follow-up is to be considered as a censored observation. Relapse or disease progression could be considered (depending on the disease being studied) as further competing events: this must be discussed with the responsible physician.

Acute Graft-versus-Host Disease (aGvHD)

The available information in the EBMT data regard the date of onset and the maximum grade of aGvHD. It is therefore possible to estimate the probability of aGvHD in a competing risks setting (death is a competing event; whether relapse/progression is a competing event must be discussed with the physician). By definition, patients alive (relapse/progression-free?) at day 100 without having experienced aGvHD are censored.

If the dates of onset are missing for the majority of patients, the analysis can focus only on the occurrence of aGvHD, which is analyzed by a logistic regression model (section 6.1). This method would however be incorrect if there is a (non negligible) percentage of censored observations or if competing events occurred before day 100.

Chronic Graft-versus-Host Disease (cGvHD)

When possible, that is if information on the date of first occurrence of cGvHD is available, it should be analyzed as a time-to-event outcome, being death (and possibly relapse/progression: ask the responsible physician) the competing event(s); data are censored for patients alive (relapse/progression-free) without episodes of chronic GvHD at last follow-up. Since chronic GvHD is defined only for patients surviving at least 100 days, the survival model should consider a left truncation at 100 days; alternatively, the time of occurrence of cGvHD must be computed from 100 days.

If information on the timing of cGvHD is not available, the outcome considered is the occurrence (eventually taking into account the grade), and the statistical model to be used is the logistic regression. Only patients surviving at least 100 days are considered to be at risk of developing chronic GvHD, therefore the analysis must be restricted to these patients. This analysis is of course not satisfactory because it doesn't take into account the occurrence of death and censoring.

Summary of the criteria to define the outcomes in survival analysis

<i>Outcome</i>	<i>Relevant event</i>	<i>Censored cases</i>	<i>Competing events</i>	<i>Population</i>
<i>Overall survival</i>	Death regardless of cause	Patients alive at last contact	-	All patients
<i>Disease-free survival</i>	Relapse of the disease or death regardless of cause	Patients alive in continuous complete remission at last contact	-	Patients in CR at transplant OR: Patients achieving CR (see note 1)
<i>Relapse incidence</i>	Relapse	Patients alive without relapse at last contact	Death without evidence of relapse	All patients OR: Patients in CR (see note 1)
<i>Non relapse mortality</i>	Death without previous relapse	Patients alive without relapse at last contact	Relapse	All patients
<i>Progression-free survival</i>	Progression of the disease or death regardless of cause	Patient alive with no indication of disease progression at last contact	-	All patients
<i>Engraftment</i>	Persistent blood cells count above the predefined level	Patients alive with no recovery at last follow-up	Death before recovery (and relapse or progression?)	All patients
<i>Acute GvHD</i>	aGvHD	Patients alive with no occurrence of aGvHD at 100 days	Death without aGvHD within 100 days (and relapse or progression?)	All patients
<i>Chronic GvHD</i>	cGvHD	Patients alive with no episode of cGvHD at last follow-up	Death without cGvHD (and relapse or progression?)	Patients surviving 100 days (set the start of the clock at 100d)

3. Variable selection, checks and coding

The third step of statistical analysis is variable selection, that is the choice of the variables which will be used in the analysis. The set of variables is composed of those describing the population plus the variables describing the course of the disease after transplantation.

In this preliminary phase of the definition of the sample, data must be checked for inconsistencies, and the number of missing cases must be calculated and taken into account. The necessary descriptive procedures to carry out these steps are summarized in section 4.

The **data must be checked** for the presence of non-sense values (e.g. negative for time to transplantation), non-consistent values (e.g. time to relapse greater than survival time) or outliers, that is, extreme values. If possible, non-sense and non-consistent values for the outcome variables should be corrected, otherwise these patients should be eliminated from the sample; for other variables, they can be set as missing. Outliers could in some case be due to typing errors, and therefore treated as non-sense values, but, if there is not enough evidence for this, they deserve particular attention (there are methods to detect their influence on the estimation of multivariate regression models, as mentioned in section 6.4).

As regards the presence of **missing values**, it is important to approach the problem in this phase of the analysis in order to finally select the sample to be analyzed³. For each variable, if the number of missing is low (e.g. less than 5% of the total number of patients), the cases with missing values can be excluded from the analysis; it is important anyway to consider whether the number of patients left at the end of this process (which must be applied for every variable of interest) still fits the goal of the study. If it is not the case, or if the number of missing is high and it is not possible to recover information from the centers, the missing cases can be properly included in the analysis as a separate category⁴, and their influence on the outcomes evaluated. In some case, it is also useful to analyze the presence and nature of missing in the data (in relation to other variables, to the centers, or to the calendar year, for example).

Finally, the **coding** of the variables has to be defined.

Discrete variables

For discrete variables, such as donor type, the proper coding can be chosen on the basis of the clinical importance of certain categories; in some cases, it is necessary to combine categories from several variables, for example to define the treatment. It is also important to collapse categories when the number of patients in a category is low.

³ The aim is to produce all the results for the same population. In fact, the results of each variable description refer only to the patients with known value for that variable, while in the multivariate regression all cases with at least one missing value in one of the variables considered are excluded from the analysis.

⁴ For continuous variables, they can be given a reference value (e.g. 0 for non-negative variables), and a variable indicating the missing values (=1 if missing, 0 if known) must be created and added to the models.

Defining intervals for continuous variables: cut points

If a continuous variable, age for example, is included as it is in a regression model, the implicit assumption is that the effect of age is linear, that is, constant for each additional year of age. This assumption should be verified; anyway, it is very common to include these variables as discrete variables, categorizing them. The choice of cut-points could be driven by clinical considerations. Otherwise, the simplest way to categorize the continuous variables is to create a certain number of groups with approximately the same number of patients, e.g. 5 intervals, each containing roughly 20% of the patients (the cut-points are the so-called quantiles).

Further transformations of the variables can be needed in the following phases of the analysis. Other indications for the treatment of variables (in particular to be used in the regression modeling) will be given in section 5.

4. Sample description

This phase of the analysis produces a description of the data with respect to the prognostic factors and the outcomes of interest.

Each variable is first described separately, in order to summarize its distribution. The type of descriptive statistics will be chosen according to the type of the variable :

Type of variable	Example	Statistics and results
Quantitative	Age	Mean and standard deviation; median and other quantiles; minimum and maximum value
Ordered qualitative	Age in classes; Grade of aGvHD (I-IV)	Median and other quantiles; frequency table
Qualitative	Gender mismatch, Donor type	Frequency table
Censored Time to event (survival analysis)	OS, DFS, PFS	Kaplan-Meier curve (with confidence interval), median (time t such that $S(t)=0.5$), estimated probability of survival at x months
Censored Time to event with competing risks	RI, NRM	Cumulative Incidence curve ⁵ (with confidence interval), median (time t such that $I(t)=0.5$), estimated cumulative incidence at x months

The relations between prognostic variables (at least, the relation between the main factors of the study and the others) are also evaluated in this phase. Suppose that the study has a main factor of interest which is “treatment”, with K possible treatments. The aim is to evaluate the differences in the K groups with respect to the other variables. Also in this case the statistical procedures depend on the kind of variables involved:

⁵ See section B7.

Type of variable	Example	Procedures and results ⁶
Quantitative with Normal distribution	Bell-shaped distribution (apply a test for this hypothesis)	Summary: mean in the K groups. T-test (K=2) or Analysis of Variance (ANOVA) (K>2)
Quantitative with non-Normal distribution	Variable assuming only positive values, positively skewed distribution	Summary: median in the K groups. Mann-Whitney Test (K=2) or Kruskal-Wallis Test (K>2)
Qualitative		Summary: cross-tabulation. Chi-Squared Test, or Fisher Exact Test when the Chi-Squared is not valid

5. Outcomes description: univariate statistical analysis

In this phase, univariate analyses are used for investigating the relationships between the outcome measures (times to certain events) and each of the potential risk factors under consideration. This provides:

- the description of the outcomes in terms of median, probabilities and survival or incidence curves, for each group determined according to the prognostic factor values.
- the preliminary step for selecting the set of variables to include in the multivariate analysis, when the number of initial potential factors is large.

The procedures depend on the type of outcome (details were illustrated in section 2):

⁶ For K=2, T-test and ANOVA give the same result, as well as M-W Test and K-W Test. The latter two are non-parametric tests (in SAS : proc NPAR1WAY ; in SPSS, NPAR TESTS). In case of paired-samples, the suggested procedures must be replaced by (following the same order): Paired T-Test, Repeated Measure ANOVA, Wilcoxon matched-pairs signed-ranks Test, Friedman 2-way ANOVA.

Type of variable	Procedure	Descriptive Statistics for each group	Test for the comparison of groups ⁷
Time to event (survival analysis) (e.g. OS, DFS, PFS)	Kaplan-Meier curve estimation	Survival curve (with confidence interval); median; survival probabilities after x months	Log-Rank Test
	(Univariate) Cox model	(Un-adjusted) Hazard ratio	LRT/Wald Test/Score Test
Time to event with competing risks (e.g. RI, NRM)	Cumulative Incidence curve estimation	Incidence curve; median; estimated cumulative incidence at x months	Gray Test
	(Univariate) Cox model (with the proper censoring)	(Un-adjusted) Hazard ratio	Log-Rank Test; LRT/Wald Test/Score Test

As it was pointed out in section 3, prognostic variables need sometimes to be properly treated before the inclusion in the Cox models.

Discrete variables: if it is not differently specified in the data-set (as a property of the variable), each variable is considered to be continuous, and, in a Cox model, its effect is estimated by a unique hazard ratio (HR) which gives the increment of hazard corresponding to a increment equal to 1 in the value of the covariate (linear effect). In the case of a binary covariate, with 2 possible (contiguous) values (such as 0-1 or 1-2), the HR estimated in the Cox model refers to the comparison of the second category to the first one (assumed to be the baseline, with HR=1), and no further action is needed. When the discrete variable has more than two categories (say K) instead it is necessary to specify that it has to be treated as “a factor”, which means that each value corresponds to a separate group. In a Cox model with a “factor” included, a global test for its inclusion is provided, as well as estimates of K-1 hazard ratios (and p-values) for the comparison of each group with the baseline group (usually the first or the last category, if not differently specified). It is usually possible to change the type of variable to “factor” in the data set, or to specify it in the estimation routine. Alternatively, K-1 indicator variables for each group (except one,

⁷ In the case of survival times, Log Rank Test and any of the tests performed in the univariate Cox model give the same result. See section B6.3 for a brief introduction to LRT, Wald Test and Score Test. The Fine-Gray test can be performed for example in S-Plus and R, using the cuminc routine (section B7).

which will be the baseline) have to be created⁸ and included all together in the Cox model; this will provide the HR for each group, while the global test can be performed looking at the Likelihood Ratio Test comparing the new model, with all the K-1 indicator variables, to the previous model, with none of them.

Continuous variables: sometimes, there are no clinically meaningful cut points for a continuous covariate, and, on the other hand, leaving it as a continuous covariate preserves the power of the model. When it is included as it is, the implicit assumption is that it acts linearly on the outcome; different assumptions can be easily made applying a transformation to the variable (e.g. logarithm, power, etc), and recent sophisticated methods, such as fractional polynomials, are able to detect more appropriate complex functional forms. As an alternative, it is possible to create a categorized version of the continuous variable and test whether the model with the continuous variable alone can be improved significantly by adding the categorized version to the model (indicating therefore a departure from the hypothesis of linearity).

6. Outcomes analysis: multivariate regression

6.1 Introduction

The aim of the multivariate analysis is to investigate the relationship between an outcome variable ('dependent' variable) and a series of other variables ('explicative' or 'independent' variables) considered simultaneously. For example, in a study for the comparison of treatments, it is important to adjust for the main risk factors, while the investigation of the prognostic factors more related to the outcomes is always carried out in a multivariate framework. This kind of multivariate analysis is also called 'regression modeling'.

Generally speaking, a statistical model is used to describe the dependence of the probability distribution of the outcome on the explicative variables. The model is specified except a series of parameters, which have to be estimated on the basis of the data; the criterion is (usually) the maximization of the **likelihood function**, which expresses the probability of the observed data for every possible set of parameters. In other terms, the estimated model is the one which assigns the highest probability of being observed to the data actually observed.

The best known regression technique is the **multiple linear regression**, which is used if the dependent variable is continuous, non-censored and Normally distributed. This

⁸ Example:

Variable: X="donor type", coded as 0-1-2 representing "Hla-id sib", "Other Rel" and "Unrelated".

Baseline category: X=0="Hla-id sib";

Indicator variables: $Z_1 = \begin{cases} 1 & \text{if } X = 1 \\ 0 & \text{else} \end{cases}$ for "Other Rel", $Z_2 = \begin{cases} 1 & \text{if } X = 2 \\ 0 & \text{else} \end{cases}$ for

"Unrelated".

case is hardly ever present in the context of BMT, where instead the main statistical tools are:

- **Logistic regression:** to be used if the outcome variable is discrete, most often binary, such as chronic GvHD (when it is not possible to study it as time-to-event);
- **Survival modeling** (in particular: **Cox model**): to be used if the outcome variable is a censored time-to-event, as OS or DFS.
- **Methods for competing risks:** e.g. for the analysis of RI and NRM.

Regression techniques present a series of common characteristics. In the following sections, we will briefly illustrate the common characteristics of the process of estimation of a multivariate model (initial variable selection, model identification and model validation). Finally, we will focus on multivariate survival models, and in particular on the Cox model, which is the most widely used model in survival analysis. In section 7 we will briefly illustrate the correct approach to a competing risks analysis, including the case of multivariate analysis.

6.2 Initial variable selection

In this phase, an initial set of potential prognostic factors for the outcome of interest must be selected. The first criterion is to base the choice on clinical *a priori* knowledge and hypotheses. The results of the previous exploratory phase (illustrated for survival analysis in section 5) can also be used: it is advisable to select the variables most significantly linked to the outcome in univariate analysis. A less restrictive p-value threshold with respect to the usual one of 0.05 is recommended, for example 0.10 or 0.20.

6.3 Model identification

In general, not all the significant variables in the univariate analysis are found significant in the multivariate analysis. This is explained by the relationships between the factors and between the factors and the dependent variable of interest. Model identification is thus essentially a variable selection process.

The purely statistical method is to use an automatic process ('stepwise' regression), which can be a) 'forward': the variables are added successively (the most significant at each step) until no variable adds significant information; b) 'backward': all variables are included in the first model and then successively removed (the least significant at each step) until the loss of information becomes significant; c) any combination of forward and backward methods. At each step, parameters are estimated according to the maximization of the likelihood function; three kind of tests derived from the likelihood are available: the **Likelihood Ratio Test (LRT)**, the **Score Test** and the **Wald Test**. Generally speaking, the LRT is to be preferred, though they are roughly equivalent, and they should reach the same conclusions. As regards the p-value thresholds for the inclusion and exclusion of the variables, we would suggest 0.05 for the 'forward' steps and 0.10 for the 'backward' steps.

It is advisable to integrate this statistical process with criteria based on clinical issues, for example always including the main prognostic factors, or the calendar year (plus of course the main factor of interest, e.g. treatment). Technically speaking, if the initial set of variables is not too large, backward selection is preferable to forward selection. It is also important, after the selection of a "final" model, to verify that

variables previously excluded from the model actually do not add a significant contribution to the model.

At this point, a “final” model, based on the effects of a set of variables, has been selected. It is now necessary to check for the presence of **interactions** between variables, which occur when the effect of one variable is different according to the level of another variable. It is recommended to check at least the interactions between the main factor of interest and the others. The significant interaction terms should be included in the model. In some cases, when the clinical relevance is negligible, they could be dropped from the model. In any case, every interaction term should be assessed from a clinical point of view.

Technically, the calculation of interaction variables is usually automatic when they are included in a model (e.g. in SPSS), but sometimes they must be explicitly calculated (this is straightforward, for example the interaction between X with values 0 and 1 and another variable Y is just the product of the two, $Z=X*Y$); it is important to include always the main terms of the variables together with their interaction (that is: the model must contain $X+Y+Z$, and not for example only $X+Z$).

6.4 Model validation

Model validation is aimed at checking that the assumptions on which the model is based are not too far from the structure of the data actually observed. Many validation methods exist, depending on the aspect of the model to verify, and on the type of regression modeling. They are mostly based on the analysis of residuals, quantities that (generally speaking) express somehow the distance between the observed data and the predictions obtained from the model. In this phase, sometimes also the impact of each observation on the estimation of the model is evaluated (“influence analysis”). An extensive illustration of the methods of validation goes beyond the scope of this document. Some recommendations for the survival regression modeling will be given in section 6.6.

6.5 Multivariate survival analysis: choice of the model

In survival analysis, we are interested in the estimation of the hazard function, which describes the instantaneous probability of occurrence of the event of interest (death, for OS; death or relapse, for DFS). Two kinds of statistical models can be used for multivariate survival analysis: the parametric models and the semi-parametric ones.

The main difference is that the first completely specify the probability distribution (or, in other terms, the hazard function) of the survival time except a small number of parameters, while the semi-parametric models avoid making assumptions on the functional form of the hazard function, and focus on the estimation of the parameters regarding the relation between the outcome and the explicative variables. This higher degree of flexibility explains why parametric models, such as the Exponential or the Weibull model, are currently less used than semi-parametric ones in biostatistics, although they have nice features, such as a higher prediction validity.

The most important semi-parametric model, and widely used in survival analysis, is the so-called proportional hazard (PH) model, or **Cox model**. Its principal assumption (“proportionality of hazards”) is that the relative rate of dying in one subgroup compared to a “baseline” subgroup is constant over calendar time. In other words if at time 0 one type of patients has twice as high a risk of dying as another type of

patients, this relative risk (hazard ratio, HR) is 2 at each point in time, from the beginning to the end of the survival curve. The estimation of the Cox model returns the HR for each subgroup considered in the analysis. These estimates can be combined with a non-parametric estimate of the baseline hazard to obtain survival curves.

6.6 Additional features and recommendations for the Cox model

The proportional hazards model has two important features that allow extensions: **stratification** and the use of **time-dependent covariates**.

Stratification corresponds to the possibility of assuming that the baseline hazard is different for sub-groups of patients, identified by a certain categorical variable, for example the kind of disease. The effects of the other covariates are assumed to be the same in each stratum. Notice that when the variable is included as a stratification variable, no estimation of its effect on the outcome is produced by the model. It is also possible to test for the difference of the effect of the other covariates in the different strata, introducing the proper interaction terms.

Time-dependent covariates are, generally speaking, variables whose value changes with time. The use of this kind of covariates is technically possible in every software, though in some cases some programming skills may be required (the syntax for SAS and SPSS is easy, but for example in S-Plus the data-set has to be properly manipulated). The interpretation of a model containing time-dependent covariates needs caution.

Let's consider the following example: in a model for survival, we want to include the effect of acute GvHD. The GvHD can only occur after transplantation, therefore the covariate X (for a patient who experiences aGvHD) is 0 from the start of the clock (date of transplantation) to the date of aGvHD, then it switches to 1. The model for survival including X will return a HR for X which evaluates the change in the hazard of death at the occurrence of the event aGvHD.

Consider now the case when the event that we want to take into account is a second transplantation; a time dependent covariate X is created exactly as in the previous case (equal to 0 until the time of 2nd transplantation, and equal to 1 thereafter). An HR of X equal to 0.9 means that the hazard of the patients who do a 2nd transplant is reduced by 10%, but it doesn't mean that the 2nd transplantation has the effect of reducing the risk of death by 10%. In fact, it could be that only patients with good prognosis get a second transplantation, so they would have had a lower hazard, even without the 2nd transplantation; it could even be the case that transplantation has actually increased their risk of dying.

This second example enlightens the importance to interpret properly the results from a Cox model with time-dependent covariates. While there is no problem when the covariate changes due to reasons not related to the process of the disease (for example, the cure center changes because the patient moves to another town due to his/her job), caution is needed when the covariate represents the occurrence of an event that is itself an outcome of the process. Generally speaking, for a correct interpretation we need also to investigate the process that determines the occurrence of each intermediate event included in the major model.

Recent developments of the so-called "**multi-state modeling**" have shown that the Cox model can be used as major tool to simultaneously model intermediate and final

events of the course of the disease. In this framework is therefore possible to estimate the probabilities of any pattern of development of the disease, taking into account the history observed up to the time point when the prediction is made. The implementation and interpretation of this kind of models are complicated, and will be illustrated in a separate document. Some readings are suggested in section D.

Model validation in Cox regression

The process of model validation in the case of Cox regression relates in particular to the assumption of proportional hazards. An easy way to check the assumption that the effect of a variable X is constant in time is to create a new time-dependent variable to represent the interaction of X with the time, such as $X * \log(t)$ or $X * t$ (notice: this is not equivalent to adding an interaction term $Z=X*Y$ as seen before, and may require some programming depending on the software you use), to add it to the model, and test if this new variable adds significant information. If it is the case, it means that the proportional hazard assumption is not satisfied, and the model has to be adjusted for this. If X is not the main factor of the study, it is advisable to use it as a stratification variable. Otherwise, a series of time-dependent covariates can be created, to represent for example early and late effect⁹.

7. Methods for competing risks

When the outcome of interest is possibly unobserved due to the occurrence of a competing event, as in the case of relapse, which is not observed if the patient dies in remission (NRM), the standard methods of survival analysis are no longer adequate. The rationale and illustration of this statement are beyond the scope of this document, anyway the key points can be intuitively understood considering that in situations characterized by competing events, two different clinical perspectives require the analysis to focus on different objects of interest:

We have already pointed out (section 2) that in the case of relapse, for example, one object of interest is the probability of having had a relapse before time t (cumulative incidence of relapse); this probability informs about the percentage of patients that will experience a relapse within a certain period after transplantation, taking into account the alternative outcome, non-relapse mortality (the focus is on *prediction*). Alternatively, or in addition to this information, we can be interested in the estimation of the instantaneous hazard of relapse at time t , which expresses the *force* of occurrence of relapse among the patients alive relapse-free at t months, that is, *conditional* on not having experienced any negative event up to time t .

In this second case, we can apply methods for the analysis of the hazard functions (as the Log-Rank test for the - crude - comparison among groups, and the multivariate Cox regression to estimate the adjusted effect of several prognostic factors), considering death in remission as censoring event.

⁹ For example : $Z_1 = \begin{cases} 1 & \text{if time} \leq \tau \\ 0 & \text{else} \end{cases}$ for “early effect”, $Z_2 = \begin{cases} 1 & \text{if time} > \tau \\ 0 & \text{else} \end{cases}$ for “late

effect”. The threshold τ can be selected to be the one to which corresponds the model with highest likelihood.

In the first case, although it is common practice to estimate the (crude) relapse incidence curve using the “OMS” curve, that is 1 minus the Kaplan-Meier curve estimated considering relapse as failure event and death in remission as censoring, this procedure overestimates the probability of relapse. The correct approach is to use the appropriate **Cumulative Incidence estimator**, though unfortunately it is not a standard tool available in every statistical software for survival analysis. In SAS, a macro is available; in S-Plus and R, it is included in the package “cmprsk” by Gray¹⁰ – command “cuminc”, which also provides a specific **test by Gray** for the comparison of cumulative incidence curves. The software NCSS has a procedure ready and easy to use.

For the multivariate analysis of the cumulative incidence curve, **Fine and Gray** (1999)¹¹ proposed a **semi-parametric model** called “proportional hazard model for the sub-distribution of competing risks”. The model can be fitted in S-Plus 2000 or in R using the command “crr” included in the package “cmprsk”; it allows stratification and time-dependent effects, but not the use of time-dependent variables. An alternative, which is currently being explored by statistical research, is to combine results from the estimation of Cox models, in a multi-state framework, to obtain estimates of the cumulative incidence; although this appears to be a more reliable method, it is still quite complicated for a standard use.

8. Matched studies

In a study for the effect of a main factor ($X=A$ or B) where a confounding factor (Y) is present (that is, Y is a variable correlated with X which has also an effect on the outcome), it is necessary to include the variable Y in the statistical analysis, performing a multivariate analysis (we call this “adjustment”).

In some cases, the two groups A and B could be characterized by a very different distribution of the variable Y ; for example, suppose that in a study patients given a reduced intensity conditioning (RIC) have a higher median age than patients with a standard conditioning. If the difference is big, the two groups are hardly comparable. In an extreme situation, all “elderly” patients receive a RIC, and all “young” patients have a standard conditioning regimen: it is therefore impossible to separate the effect of the type of conditioning from the effect of age.

¹⁰ The package « cmprsk » is easily available in R, a complete statistical package very similar to S-Plus, which is distributed freely on the web (<http://www.r-project.org/>).

Once R is running, and having a connection to the internet, to install “cmprsk” it is sufficient to execute, the first time, the commands:

```
> options(CRAN= "http://cran.at.r-project.org")
> install.packages("cmprsk")
```

To use the methods for competing risks, it will then be sufficient to execute, at the beginning of each session:

```
> library(cmprsk)
```

R has also a library “survival” that performs all the analyses discussed in this report (Kaplan-Meier method, Cox regression, parametric models, but also random effects survival models, and others). It is possible to download pdf help files for these libraries.

¹¹ The reference paper is : Fine JP - Gray RJ on *Journal of the American Stat. Assoc.*, 1999, 94: 496-509. For the Gray test : Gray RJ on *Annals of Statistics*, 1988, 16, 3 : 1141-1154.

In these situations, it is possible to plan a matched study, that is, to select a comparison series (patients treated with standard regimen) that is identical, or nearly so, to the index series (RIC patient) with respect to one or more potentially confounding factors (age, in our example). The mechanism of the matching may be performed subject by subject, which is described as individual matching, or for groups of subjects, which is described as frequency matching. The general principles that apply to matched data are identical for individually matched or frequency matched data.

Matching also occurs in randomized studies where a balanced randomization has been applied.

In any case, whenever the study is planned as a matched study, the stratification variables must be systematically included in the models to be estimated (*conditional* logistic regression and *stratified* Cox regression should be used).

Notice that if matching is not necessary, the statistical analysis is less efficient; another drawback is that the effects of the stratification variables cannot be estimated. Therefore, we recommend when considering a matched study to discuss the plan with a statistician.

9. Interpretation of the results

When data are collected from the current clinical experience without any experimental research protocol, as in the case of the EBMT registry, it is important to correctly interpret the results of the statistical analysis. In particular, it is fundamental to stress that in principle it is not possible to infer any causal relationship from the assessment of statistically significant associations.

So for instance, if the statistical analysis shows that the overall survival is significantly different for two therapeutic regimens, this does not prove that one of them is better than the other. The information should be considered as having interest in itself as a description of the outcomes, for patients and clinicians, but also to generate hypotheses for future research, and in particular to elaborate a plan for a randomized clinical trial, which is the correct method for comparing treatments.

C Presentation of the results

1. Kaplan-Meier and Cumulative Incidence curves

We recommend to report graphically the survival probability and the probability of occurrence of an event in time as a curve plot.

The caption of the plot should indicate the selection criteria applied to estimate the curve.

The size of the patient population must be indicated by one of the following methods:

- Tick marks to indicate individual patients who have not yet reached the critical event
- Numbers specifying the number of patients at risk at various time intervals
- Confidence intervals (usually 95%) plotted as error bars in the figure or specified in the legend below

The scale for the ordinate should be the probability (0.0-1.0) or the percentage (0-100%).

The scale for the abscissa should be days, months (in numbers divisible by 12) or years.

The plot of the survival (or cumulative incidence) curve should be terminated when less than five patients are at risk

As examples of figures, consider the following:

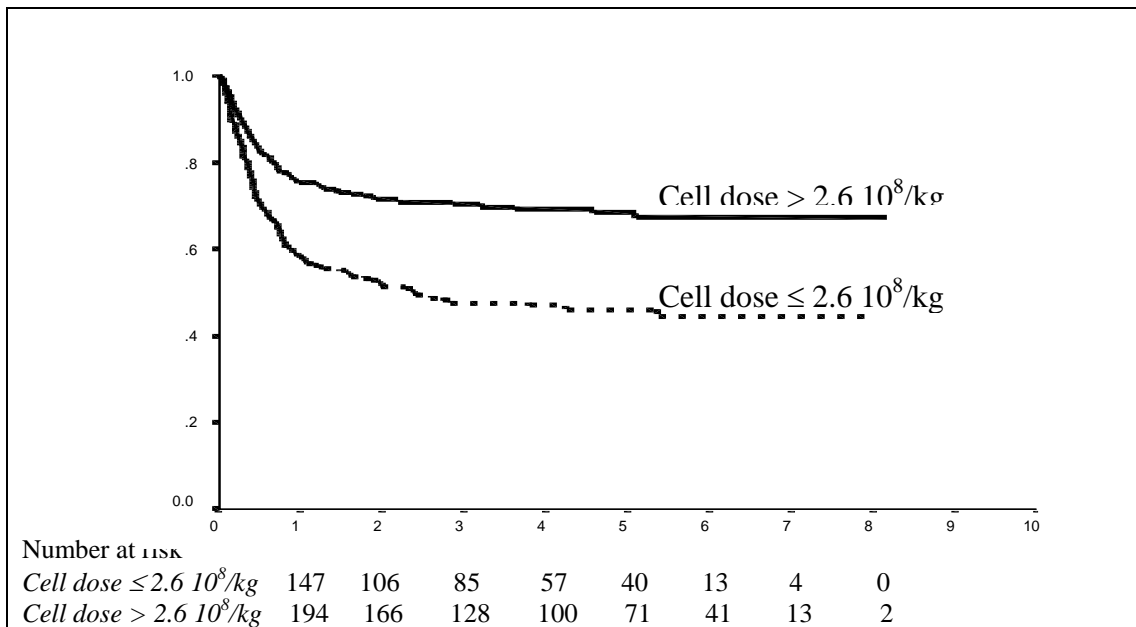


Fig 1 : Overall Survival by cell dose infused in AML CR1.

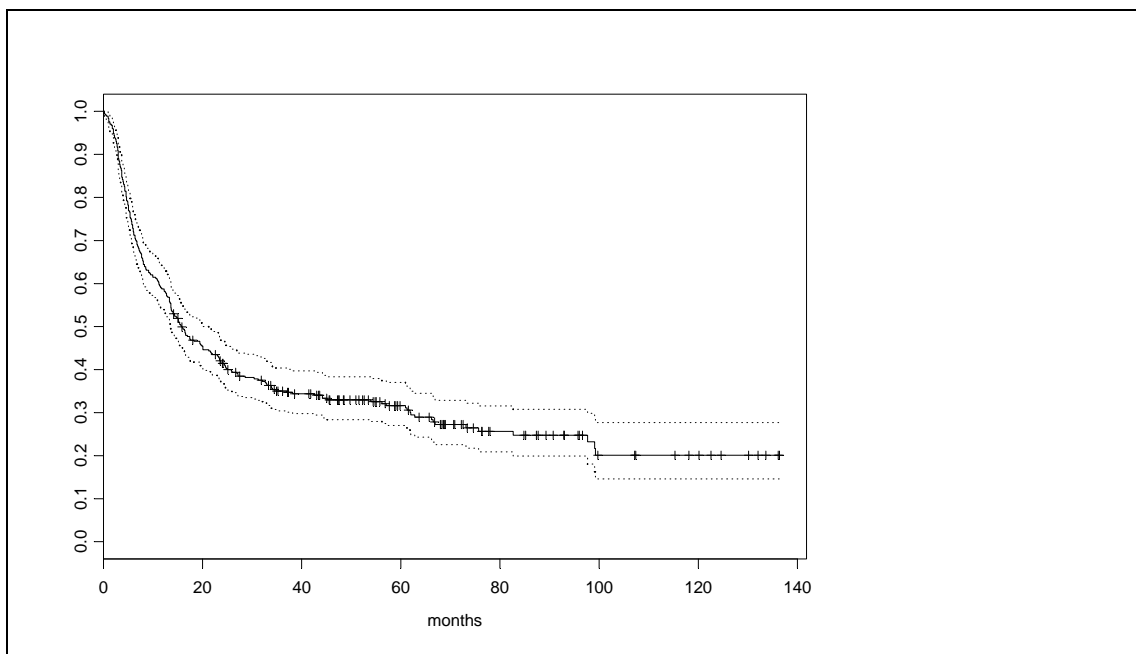


Fig 2 : Survival after Relapse (with 95% Confidence Interval).

As a further report, a table of survival probabilities and cumulative incidence estimates at x years can be included. The results can be expressed as percentages. For example:

Table 1: HLA-id sibling recipients

<i>SURVIVAL and DFS KM estimated probabilities and RELAPSE and NRM cumulative incidence estimates at 3 years</i>					
	N 3 yrs	SURV (%)	LFS (%)	REL (%)	NRM (%)
total	100	71	59	21	20
stage					
CP1	93	75	63	17	20
no CP1	7	46	35	49	16
Gender mismatch					
male-female	25	71	62	14	24
other combinations	75	71	58	24	18

N 3 yrs: number of patients at risk at 3 years.

2. Cox models

The list of the covariates initially included in the model and the criteria for the selection must be specified in the statistical methods section. A table with the estimated Hazard Ratios, 95% confidence intervals and p-values at least for each covariate significantly associated with the outcome must be presented. In case of interaction terms, the main effects of the covariates involved in the interaction must be present in the table, even if the main terms are not significant.

As an example, consider the following:

Table 2: Multivariate analysis for Overall survival

Riskfactor		HR	95% CI	p-value
Stage	CP1	1		
	AP	2.0	1.3-3.0	<0.01
Mtx	No	1		
	Yes	0.6	0.4-0.9	0.038

“Adjusted HR” for each riskfactor is adjusted for all other factors in the same model.

95% CI: 95% confidence interval

D Further readings

- 1) Bull K, Spiegelhalter DJ (1997): “Tutorial in biostatistics survival analysis in observational studies”, *STATISTICS IN MEDICINE*, 16, 1041-1074
- 2) Hosmer DW, Lemeshow S (1999): *Applied survival analysis*, Wiley ed.
- 3) Klein JP, Rizzo JD, Zhang MJ, Keiding N (2001): “Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part 1: Unadjusted analysis”, *BONE MARROW TRANSPL*, 28, 10, 909-915
- 4) Klein JP, Rizzo JD, Zhang MJ, Keiding N (2001): “Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part 2: Regression modeling”, *BONE MARROW TRANSPL*, 28, 11, 1001-1011

For competing risks analysis:

- 5) Gooley TA, Leisenring W, Crowley J, Storer BE (1999): “Estimation of failure probabilities in the presence of competing risks: new representation of old estimators”, *STATISTICS IN MEDICINE*, 18, 695-706

For multi-state models:

- 6) Keiding N, Klein JP, Horowitz MM (2001): "Multi-state models and outcome prediction in bone marrow transplantation", *STATISTICS IN MEDICINE*, 20, 12, 1871-1885
- 7) Klein JP, Keiding N, Shu YY et al (2000): "Summary curves for patients transplanted for chronic myeloid leukemia salvaged by a donor lymphocyte infusion: the current leukemia-free survival curve", *BRIT J HAEMATOL*, 109, 1, 148-152