Bone Marrow Transplantation (2013) 48, S1–S37
© 2013 Macmillan Publishers Limited All rights reserved 0268-3369/13
www.nature.com/bmt

EBMT STATISTICAL GUIDELINES

Suggestions on the use of statistical methodologies in studies of the European Group for Blood and Marrow Transplantation

Simona Iacobelli, On behalf of the EBMT Statistical Committee

European Group for Blood and Marrow Transplantation & Università di Roma Tor Vergata, Italy

This document explains some of the relevant methodological issues involved in planning a clinical study using survival and time-to-event outcome data, particularly in the field of haematopoietic stem cell transplantation, and indicates the appropriate statistical methods to use for the analysis. As the majority of these methods are commonly used in survival and event-history analysis, this document discusses their potential and limitations with reference to common SCT research situations. Some reference is given to methods, recently appearing in the literature that may be capable of handling complex investigations. These guidelines also address various practical issues, such as recoding or transforming variables in regression models or reporting results.

Bone Marrow Transplantation (2013) **48,** S1–S37; doi:10.1038/bmt.2012.282

Keywords: statistical methods; survival analysis; competing risks; statistical guidelines; stem cell transplantation research

Target audience: The Statistical Guidelines are a reference for all studies of the European group for Blood and Marrow Transplantation (EBMT). Additionally, they may be useful to a more general community of researchers in stem cell transplantation (SCT), or in any field requiring the analysis of time-to-event data, as well as to anyone involved in clinical research, including statisticians, statistical analysts and clinical investigators. It is assumed that the reader has some elementary understanding of statistics and/or experience in clinical research.

Suggestions on how to use this document

This document is suitable for statisticians at all levels of expertise. However, these Guidelines do not replace textbooks addressing survival analysis or the extensive literature on traditional and more recent methods of analysis (reading suggestions are provided throughout the document). These Guidelines are intended to present one view on interpreting the most common methodological issues encountered in SCT research, and to promote a unified approach for EBMT studies.

Clinical investigators involved in study planning or who are more generally interested in methodological issues might wish to restrict themselves to Chapters 1 and 2. Those sections marked with asterisks (*) address more complicated issues, and may be initially skipped. Additional topics of general interest for planning and interpreting studies are addressed in those sections of the remaining chapters marked with a circle (°).

In our experience, all professionals involved in the acquisition and reporting of data and in the publication phase of research are interested in the general ideas and concepts contained in this document. We therefore hope they will find the Guidelines as a whole useful, particularly Chapter 1, Sections 2.1 and 2.3, and Chapter 3, which are dedicated to illustrating the most common methods for descriptive analysis, preliminary study phases and the reporting of results.

To support non-statisticians in performing these analyses in practice, the Appendix provides a list of commands and procedures for implementing the basic statistical methods in R, SAS and SPSS.

On a final note, although some parts of this document are rather technical, we avoid almost entirely any discussion of theoretical issues and make an effort to present all arguments in a manner that is accessible to readers of all backgrounds. Some mathematical formulas are given; these are not essential (and can thus be skipped), but they might be useful for comprehension and are presented within the context of purely intuitive remarks.

1. General and introductory issues

1.1. Objects of interest in SCT research

Haematopoietic SCT is used for the treatment of several different diseases, and the object of interest in any EBMT study may therefore vary depending on context. Generally speaking, of greatest interest is the relationship between

Statistical Committee: Myriam Labopin (Chair), Ariane Boumendil, Ronald Brand, Ulrike Poetschger, Antonella Santucci, Stefan Suciu, Richard Szydlo, Liesbeth de Wreede.

Sponsorship Statement: This supplement is sponsored by the European Group for Blood and Marrow Transplantation (EBMT).

patient characteristics or treatments and a clinical outcome, usually the occurrence of an event and the time between transplantation and when the event occurs. Apart from outcomes that are treatment- or disease- specific, there are several common objects of interest that will be presented in Chapter 2, including overall survival (OS), relapse incidence and relapse-free survival (RFS). In this section, we introduce some of the concepts (and terminology) in very general and intuitive terms.

The duration of survival is an object of interest in most transplant studies. In Chapter 2, we will see that even when the object of interest is clear, a formal definition requires us to specify the time of 'origin' and the time the clock stops, as well as to create an indicator to determine whether the event of interest actually occurred. In fact, the observation itself takes place during a specific time interval called follow-up; if the follow-up period is not sufficiently long, some patients may be alive on the last occasion they are seen, but their exact total survival time (the information of interest) will not be known. These are not 'missing values', however, because we actually do have some information regarding the total survival time. If, for example, the patient was last seen x months after transplantation, we know that the time-to-death value is larger than x. This type of 'incomplete' observation is termed censored, and the field of statistics that processes this type of data is called survival analysis.

Censoring and the methods of standard survival analysis do not exhaust the statistical framework that any investigator in SCT should be aware of. Often, the event of interest is not death, which sooner or later occurs in every patient, but rather events that may or may not occur at all, depending on whether some other event occurs before the one of interest. If we are interested in the recurrence of a disease in patients who are in remission at the start, we may observe recurrence in many of them with a sufficiently long follow-up, but we may also have patients who die in remission without prior recurrence. In the latter case, the object of interest, time to recurrence, is not observable. These two events, recurrence and death without prior recurrence, are competing risks. The statistical methods used to analyse the probability of the occurrence of an event with competing risks are different from those methods used for standard survival, because the type of information is different. We will see in Sections 2.1.2 and 2.2.2 that the occurrence of a competing event is not equivalent to censoring, but in fact the patient stops being at risk for the event of interest, while in the survival setting the censored patient remains at risk. Only patients who are alive at last follow-up with no events occurring ('failure-free') are censored cases in a competing risks situation.

In SCT research, we commonly encounter outcomes with competing risks, such as relapse (disease recurrence), haematopoietic recovery (engraftment), the onset of GVHD, and the achievement of CR (when patients can be transplanted with active disease, as in lymphoma or myeloma). All of these events may be precluded by the occurrence of death, and some may have additional competing events depending on the context (for obvious reasons, these Guidelines cannot be specific for each disease or each particular situation). The analysis of mortality from a specific cause is another example of a competing risks situation (dying from any cause other than the one of interest is a competing risk).

Some of these events are often, or were at least traditionally, evaluated within a fixed timeframe following transplantation. Leukocyte engraftment was evaluated within 30 days, and the case was designated to be acute GVHD when it occurred within the first 100 days. When the ending time threshold is very close to the origin, then the follow-up may be complete up to the day of evaluation, in the sense that for all patients, it is known whether the event of interest or a competing event (death and possibly others) occurred by the time of evaluation, and no patient is lost to follow-up before the fixed time. In this case, it is possible to disregard the statistical methods of survival (competing risks) analysis and use either percentages to estimate the total incidence or other simple summary statistics for the analysis (Section 2.2.3).

Finally, in a competing risks situation where the competing events all represent different causes of failure, the investigator may be interested in the total probability of failure, regardless of the cause. Here, the object of interest is a combined endpoint, the time to the first failure observed for the patient. For example, RFS is the duration of survival until the first recurrence or death. Because this combined failure will certainly occur (and would be observed during a sufficiently long follow-up), this type of endpoint can be analysed using the standard methods of survival analysis.

Survival-like events or events with competing risks that are observed over time and can be censored are the main type of outcomes of interest in SCT.[1] In Section 2.1, we will return to the definition of proper endpoints in SCT studies and their analysis with more comments and examples.

The clinical context and the goals of the study must govern the definition of the endpoints of interest, which must be relevant for study objectives, consistent with the disease and/or treatment being investigated, and feasible with respect to the data available. To define the endpoint properly, the statisticians and the clinical investigators should discuss the clinical framework at length, clarifying the role of the relevant events during the disease history.

1.2. Type of studies

In SCT, as in any field of medical research, the possibility of producing evidence for the relationships between different phenomena depends on the type of study. It is far beyond the scope of this document to illustrate the characteristics, potential and limitations of the different

[1] Other outcome variables in longitudinal studies, such as recurrent events or repeated measurements, are not common in current SCT research. Thus, we provide only one example of analysis in a situation characterised by multiple recurrences of relapse, which will be approached in the context of multi-state models (the analysis of current leukaemia-free survival). However, other statistical approaches may be more suitable in other situations. Methods for longitudinal data and for multivariate survival outcomes are not illustrated in these Guidelines. Even less common is analysing a continuous outcome variable, such as the level of WBC. Methods for the analysis of this type of outcome (*t*-test, analysis of variance, multiple regression and so on) can be found in any statistics textbook. A few are mentioned in Section 3.1.

type of studies, but extensive discussions may be found in many classical biostatistics and epidemiology textbooks. Here, we introduce the major issues for each of three main study categories. Although we find it convenient to refer to the same categories and use the same acronyms from other official EBMT documents,[2] the contents of this section nonetheless generally apply to any retrospective, observational prospective or experimental (interventional) prospective study.

- A registry-based study (RBS) uses the data available from the EBMT registry, which retrospectively collects data for each patient who receives a SCT. In RBS, all data are (in principle) available at the moment the study is initiated; in a sense, the data refer to the events that have already occurred. Thus, this type of study is called 'retrospective' with reference to planning and data collection.[3] This study category is similar to those studies based on case series collected from standard clinical practice.
- Prospective data collection can be initiated to follow the disease history of patients from the moment they are transplanted. The interest may be in patients with a particular characteristic or diagnosis, or in specific type of transplantations or treatments. We call these prospective collections (observational) as non-interventional studies (ONIS or NIS). They are more generally referred to as 'observational studies'. The fundamental difference between a clinical trial and a NIS is that a NIS does not affect the choice of treatment nor does it influence the clinical management of the patient, which follow medical decisions that are made completely independently of participation in the study (thus, the study is 'observational' in the traditional epidemiological meaning). Similar to a clinical trial, a NIS also follows a protocol that fixes the inclusion and exclusion criteria and sample size, it may indicate a schedule for the clinical assessment of the status of the patient, and it provides instructions for data collection.
- Prospective interventional studies, where treatment is determined by a protocol, are also called prospective clinical trials (PCTs). These studies are conducted in an experimental setting, and all aspects of the study, including treatment (type, dosage), the management of secondary effects, scheduling and the type of clinical evaluation, follow a prescribed protocol that maintains the possibility of changing or stopping treatment according to individualised medical decisions or the possible withdrawal of consent by the patient. Thus, in a PCT, patients are recruited based on their adherence to the inclusion and exclusion criteria, and the treatment decision depends on their participation in the study. Conversely, in a NIS, participation in the study depends on the decision to treat.

The listed order of these studies describes the increasing potential of the study to extract from the data evidence of (causal) relationships between prognostic factors or treatments and outcomes. Of course, this potential also corresponds to higher costs, longer duration and, possibly, to reduced feasibility.

The higher potential of clinical trials arises from their ability to control sources of heterogeneity, which can affect an analyst's ability to detect the presence of the relationships of interest, and from the allocation of treatments according to a protocol that facilitates the establishment of causal relationships. Strict inclusion and exclusion criteria control the biological variability of the patients. Treatments are homogenised in terms of type, dosage, schedule and so on, and, most importantly, they are assigned independently of the characteristics and current status of each particular patient. All assessment methods, as well as the management of secondary outcomes, follow the same criteria. Moreover, it is possible to control for certain types of bias through the use of proper methodological devices, such as randomisation or blinding.[4]

The main limitation of RBS is the fact that with the data collected retrospectively, we almost completely ignore the motivations behind the choice of treatment, yet it is highly likely that the choice was related to the patient's characteristics and status at the moment the decision was made. Thus, the effect of the treatment on outcome is greatly confounded by other factors. For example, we may observe that patients who received treatment A had better outcomes than patients who received treatment B, but we cannot attribute this difference to a causal relationship treatment–outcome ('treatment A is superior to treatment B') because the explanation could be that treatment A was assigned to patients who already had a better prognosis before treatment than those who received B. This limitation is only partially amendable by applying statistical methods to 'adjust' for patient characteristics; we cannot, for example, control for unknown, unmeasurable or unmeasured factors (more discussion about this problem appears in Section 1.3).

Although subject to potential bias and confounding factors, RBS can be very useful for clinical research. The large amount of available information in the EBMT registry allows researchers to conduct exploratory analyses that provide descriptions of patient characteristics and outcomes, and investigate relationships that may be useful for generating hypotheses for future research, particularly when planning prospective observational studies and clinical trials. The drawback of having large amounts of information, however, is the potential to overanalyse and overinterpret the results. It is recommended that any analysis conducted follows a plan constructed around a series of hypotheses, while remaining aware of the methodological limitations, rather than rely on 'data mining' that emphasises whatever statistically significant results can be found. With respect to the hazards of misinterpreting significant results, comments on hypothesis testing can be found in Section 1.4.

[2] EBMT internal guidelines for the conduct of studies are available for the investigators through the EBMT study offices.

[3] This terminology might create some confusion for statisticians, as the perspective is 'prospective', that is, longitudinal. In statistics and epidemiology, the term 'retrospective study' is often used with a different meaning, as for example in the case–control study.

[4] The discussion of these issues is beyond the scope of these Guidelines. EBMT adopts the ICH guidelines for the design of clinical trials.

Observational studies (NIS) partially control the heterogeneity and biases relating to population selection, the observation of outcomes and data collection, and are therefore capable of delivering stronger evidence of causal relationships than RBS. Caution, however, is still required because a NIS is limited by the fact that all decisions regarding treatment may depend on several confounding factors (patient characteristics, intermediate outcomes, the centre's attitude and so on). Thus, for a NIS, it is valid to follow a predetermined plan of analysis and avoid making a data-driven (*P*-value-driven) analysis or overinterpreting the associations detected. Similarly, prudence must be adopted for any unplanned analyses on the data collected during a PCT (such as subgroup analyses not foreseen in the study protocol), and when publishing the results, it should be specified that they are 'exploratory' analyses. This is necessary because all of the 'machinery' of the trial (eligibility criteria, controlled experimental situation, sample size, randomisation and so on) was built to examine specific questions and does not 'protect' against confounding or other forms of bias when looking at other questions.

1.3. Phases of the study

A good study plan consistent with the goals and practical feasibility of the investigation—with the ability to collect the necessary information or, in RBS, the current availability of data—is fundamental to the success of the study. Thus, the planning phase (Chapter 2) should be conducted with particular care, even if it requires considerable time and effort. The statistician is involved at several points in this phase.

First, the investigators must choose the type of study, and if the choice is a PCT, the type of experimental design.[5] Closely related to the choice of study type is the definition of the outcome variables, or endpoints, the statistical objects that will be used to measure the object of interest and to assess its relationship to certain factors. For example, if the main object of interest is the effect on the risk of early death of condition A compared to condition B, then the endpoint could be the OS probability at 1 year post transplant. In clinical trials (and in NIS), a distinction is made between the primary endpoint and the secondary endpoints. The sample size of the study is computed based on the required power for testing an hypothesis on the primary endpoint, or on the ability to achieve satisfactory precision of the estimates (controlling the width of the confidence intervals).

The type of disease, the treatment, and other clinical and biological issues largely determine which endpoints should be chosen. The EBMT releases documents providing criteria for correct clinical/biological definitions. These definitions must then be analysed in terms of statistical rigour, adjusting the analytical methods to align with the type of endpoint. This study phase thus requires combining both clinical and methodological issues. Section 2.1 discusses the most relevant aspects for these factors and provides a brief, insightful introduction of the most common endpoints used in SCT. It will be seen that for the statistician working on SCT, it is essential to understand the role that a series of events such as engraftment, chimerism, GVHD and relapse have on SCT to plan and conduct a proper analysis. Notice also that the growing knowledge of the underlying biology and the availability of new treatments and diagnostic methods require the statistician to periodically reconsider definitions and statistical approaches (one example is the case of GVHD, see footnote 30). Consider also the importance of knowing the exact information recorded in the database, and in the case of a data registry such as the EBMT, knowing whether a definition has changed over time. This is just one example of the fact that during a clinical study, it is vital to maintain effective interactions between the responsible physician, the statistician and the study coordinator.

During the planning phase, investigators also inventory all factors (patient characteristics, transplant types, and so on) that may have an effect on the endpoints and/or their relationship to the main objects of interest. These elements are taken into account to define the study population, guarantee a comprehensive description of the population, and plan an effective approach in controlling for bias and confounding factors.

A general definition of bias is systematic error; more specific definitions would require more specific discussions than what this document is intended to provide. We will specifically address selection bias where the study population does not represent the target population at which the study is aimed, thus generating study conclusions that cannot be generalised (see Section 2.3). We present several situations of biased comparisons throughout the document. One major issue related to bias is confounding, which arises when the main factor (for example, the subtype of disease, A or B) is associated with another factor (such as age) that is influential to the outcome (say OS). Suppose the study aims at comparing the OS between the two disease groups A and B, and the analysis shows that survival is worse in group A than in group B. However, elderly patients have worse OS, and (it does not matter whether it is natural, the result of selection criteria or just pure chance) it just so happens that the percentage of elderly patients is higher in group A than in group B. Given this scenario, we could not state that there was an association between disease subtype and OS because the difference in OS observed between groups A and B could just as well be attributable to the association between age (the confounder) and OS. This problem is ideally overcome by comparing A and B among patients of the same age.

Clinical trials can be designed to reduce bias and confounding using specific devices such as randomisation or blinding.[6] We focus here on the approaches that can be applied to any kind of clinical study. Intuitively we have shown that a comparison between subgroups should control for a potential confounder by keeping it equal among the groups. This could essentially be done in three

[5] In SCT trials, the comparison between treatments is usually made as parallel groups. In other settings, popular designs include the crossover, the factorial design and others.

[6] In PCT comparing two or more treatments, randomisation, that is, the assignment of a patient to a treatment arm following a random process and independently of the patient's characteristics, is designed to (ideally) produce groups similar in terms of known and unknown risk factors. Blinding is adopted to reduce biased assessments of clinical responses.

ways[7,8] (illustrated with reference to an example where age is the confounder):

- Restriction would involve reducing the study population to patients belonging to the same age group. The main drawback of restriction is the problem of generalising the conclusions of the study to other age groups. Additionally, if restriction is performed only at the statistical analysis stage, it implies a reduction in sample size, and thus reduced power for testing and a lower precision for estimates.
- In stratification, the population is divided into subgroups (or strata) with the same or similar value for the confounder (here, in age groups), and statistical analysis uses methods developed for stratified data; the comparison between A and B is performed separately in each age group, and the results are then combined. This approach requires that the difference in outcome between A and B is similar in all groups (that is, that there is no effect modification, or in more statistical terms, no interaction with age). This approach is preferable to restriction, provided the latter assumption holds.
- Regression analysis or other more complex statistical methods (as illustrated in Chapter 4) have the advantage with respect to stratification in controlling for multiple potential confounders simultaneously. Regression models (for example, the Cox model) provide an estimation of the 'net' effect on the outcome of each factor x in terms of the difference that could be attributed specifically to the variation of factor x alone, all other factors being fixed. This type of analysis is necessary in the presence of confounding, but it is also useful if no other prognostic factor (associated to the outcome) is associated with the main factor because controlling for all effects reduces unexplained variability and increases the power to detect significance for the main effect. Regression models are also used in studies where the interest is not limited to a single factor, but the aim is to elaborate a prognostic model.

The need to control for bias and confounding thus affects the criteria used for the selection of the population and the choice of the statistical methods used for the analysis. In RBS, some decisions can also be made after data collection, as several aspects of the study can be reviewed and refinements can be made during the preliminary data analysis phase. In a NIS and in a PCT, however, all decisions must be made during the planning phase, both in terms of the definition of the study population (fixing inclusion and exclusion criteria) and selecting the statistical analysis plan. The validity and strength of these studies actually rely on the adherence of the statistical analyses to the methods originally chosen and described in the protocol; additional analyses could be performed, but they should be considered 'exploratory' in nature, and should be interpreted similarly to evidence obtained from the non-experimental observational studies (Section 1.2).

The analysis phase (Chapter 3) begins when the data are available. This phase is illustrated mainly in the context of a study aimed at comparing subgroups defined according to a main factor of interest.[9]

A preliminary descriptive analysis is performed to confirm the quality and consistency of the data and to review and refine the study plan,[10] identifying any potential problems resulting from outliers, missing values, unexpected associations and potential confounding, and so on.

The analysis then proceeds with a full description of the observed study population, which is the first result of the study. The distributions of the main variables of interest are described separately and in association with other variables, when relevant. In particular, the subgroups of interest are compared (usually with significance tests) in terms of characteristics and outcomes. This is referred to as marginal (or univariate) analysis, which, as we have just seen, can be affected by confounding and usually by heterogeneity. Thus, a more complex adjusted comparison is usually performed (Chapter 4). In view of the key role of statistical tests in the analysis and interpretation of results, we dedicate Section 1.4 to this topic.

The final phase of the study is the presentation of results in tables and graphs (Section 3.5), and a brief description of the relevant statistical issues of the study, particularly the definitions of the endpoint and the methods of analysis used. Effective communication is, of course, important for the success of the study. The task of the statistician is to produce tables and graphs that are precise, comprehensible, informative and honest, in the sense that they should not attempt to 'hide' some limitation of the study, such as the presence of missing values, a short follow-up time or the weakness of the statistical significance of any given difference. The reader of the article should receive all of the necessary information to critically appraise the results of the study and use the derived knowledge appropriately in his/her research, as well as in clinical practice.

1.4. Remarks on statistical tests

First, it is worth recalling the general construction of a statistical hypothesis test (although only in simplified terms,

[7] A further 'traditional' possibility is a case-matched study, in which each patient from group A is matched to one or more patients from group B who are identical, or nearly so, with respect to one or more potential confounders. The analysis must apply specific methods for matched pairs. This method is more difficult to apply than the others, it is less efficient (using only a part of the available information) and it has less potential than regression methods for investigating the role of all possible confounders. Matching does have a function in the propensity score method (Section 4.4).

[8] Recent literature proposes other approaches to the specific problem of removing potential confounding. In the context of the comparison of two treatment groups in a non-intention to treat non-randomised study (where, for example, propensity scores are used, Section 4.4), it is worth mentioning the use of adjusted survival curves.[42] This method uses inverse probability weights to create adjusted survival curves to artificially create comparable groups in terms of all factors that could have influenced the choice of treatment.

[9] A different yet frequent goal of many studies is proposing a new risk score, or building a 'good' prognostic model for outcome prediction in general. Although this problem can also be approached using the methods described in this document (regression models), a proper investigation requires specific validation techniques whose illustration would be beyond the scope of these Guidelines. Throughout the document, we will identify issues such as this and suggest texts[28] that can be consulted that illustrate methods for assessing the predictive value of the model.

[10] As indicated, the refinements of the study population (and of the statistical analysis plan) apply only (or especially) to RBS.

and only for the case of a two-sided test for the presence of a difference), after which we will discuss the analysis and interpretation of results.

A test aimed at proving that there is a difference between two groups defines two hypotheses, the null hypothesis H_0 that states that there is no difference, and the alternative hypothesis H_1 that states that there is a difference.[11] The hypotheses do not refer to the observed data, but rather to the difference 'in nature' or in the general population from which the observed sample came. The null hypothesis represents a 'neutral' situation, one that is considered to be true until the observed data strongly indicate that H_0 does not hold. In this case, the null hypothesis is 'rejected'. A statistical test is a procedure for 'gathering support from the data against the null hypothesis'.

A statistical test is essentially a mathematical procedure based on a probability model that returns a number, a *P*-value, quantifying the probability that, even if in the theoretical, general population there is no difference (that is, H_0 is true) in the sample we observe due to pure chance the difference actually observed, or an even larger one. A very small *P*-value, then, suggests rejecting the null hypothesis because it is not supported by the data, so the smaller the *P*-value, the stronger the confidence one has that rejecting H_0 is a valid conclusion. When the *P*-value is small, the difference observed in the data is said to be 'significant'.

It may appear thus far that examining the *P*-value is a good way of answering a research question regarding the presence of a difference, but several arguments show that this is not sufficient. One argument is substantial: 'significant' means 'very unlikely due to pure chance', but it does not mean 'clinically relevant'. Another argument is technical: the *P*-value depends on the sample size, and it decreases with larger samples. When these two arguments are combined, we could have a very small *P*-value even if the difference is negligible in clinical terms. This simple fact is not taken into account every time the result of a test is evaluated when only the *P*-value is considered.

A third argument is related to the application of decision rules that are based on fixed significance thresholds to establish whether or not to reject H_0. According to a common practice that we will call 'the 5% rule', a $P<0.05$ is thought to 'prove' the presence of a difference, while at the same time it is not considered worth the effort to further consider any difference with a *P*-value >0.05. The 0.05 threshold is the value commonly chosen for the alpha parameter of the theoretical paradigm of statistical hypothesis testing, also known as the 'significance level' of the test, that represents the probability of a type I error, which is the probability that the decision rule leads to the rejection of H_0 when H_0 is true.[12] Thus, when we apply the 5% rule, we accept that we run a 5% risk of making this type of error. The rule is acceptable in some cases, but it is not reliable enough to be the gold standard. It is recommended that flexibility and critical capacity are adopted in the interpretation of *P*-values.

We now have the elements required to recommend the use of confidence intervals. A 95% confidence interval for a difference provides more complete information than a *P*-value. It shows the minimum and maximum difference that you can expect in the population consistent with the observed data and from which you can determine clinical relevance. Moreover, when the 95% confidence interval does not include the value corresponding to the absence of any difference, it means that the *P*-value of the two-sided test is smaller than 0.05 (that is, that the difference is significant at the 5% level). Whenever possible when reporting subgroup estimates for comparison or the effect of a factor from a regression model, confidence intervals should be reported in addition to or in place of the *P*-values.

Thus far, we have considered the case of a test on a null hypothesis, H_0, stating that there is no difference. We saw that a small *P*-value indicates that the data 'prove' (provide strong evidence of) the presence of a difference. Perhaps counter-intuitively, because of the construction of statistical tests, a large *P*-value does not mean that the data prove the absence of any difference. We can only say that the data do not provide enough evidence to reject the null hypothesis that there is no difference. If you are willing to use the data to demonstrate that there is no difference, or that there is a negligible difference, then you must apply specific tests for equivalence or non-inferiority, where the null hypothesis states that there is a difference larger than a certain value, and the alternative hypothesis states that the difference is smaller than that value.

One problematic aspect of hypothesis testing is known as the problem of multiple testing. In a statistical analysis, we usually apply a large number of tests and interpret their results independently from one another. For example, say that we draw conclusions using the 5% rule to evaluate significance. This implies that for every test, we run a 5% (alpha) risk of type I error. Consequently, the total probability of making at least one such error, that is, wrongly drawing the conclusion that a difference exists, is actually higher than the fixed 5%. In other words, the more tests we perform, the higher the chance that we will make the wrong conclusion, and the more we increase what we might call the 'false discovery rate'. The inflation of type I error can be controlled using several techniques that may be more or less easy to understand and/or implement, and more or less commonly applied and accepted in clinical research.[1] Apart from using simple approaches such as the Bonferroni–Holm correction (Section 4.2.3), it is not possible to suggest solutions for each potential situation. We can recommend avoiding the overinterpretation of significance, particularly when the *P*-values are close to the usual significance level (for example, $P=0.04$), the results are counter-intuitive, or the results are not fully consistent with other results in the same study or from other studies. It is fundamentally important to avoid 'fishing for significance' (such as by testing every possible association) or

[11] In a two-sided test, H_1 states that there is a difference of whatever sign, positive or negative. A one-sided test fixes the sign of the expected difference.

[12] Another important parameter whose meaning is worth recalling is the power (1−beta), which is the probability that the decision rule leads to 'proving' the presence of a difference that is actually present in the target population. Of course, this parameter should be high; it is, however, limited by the requirement of having a small alpha. The power is a key parameter in determining the sample size of a study.

including a continuous variable into a model by looking for a cut-point that makes it a significant risk factor. A sensible restriction to a few questions based on clinical and biological knowledge would provide a reasonable guarantee against finding 'false positive results' without using complicated statistical adjustments.

As a final remark, when investigating the presence of a difference you may, despite applying the correct statistical procedure for testing and correctly interpreting the *P*-value, still report the wrong conclusion if the study is biased or if you presume the presence of a causal relationship in a context where this relationship is not appropriate.

- **A statistical test corresponds to probabilistic reasoning. It is based on a mathematical model and adopts specific criteria for decision-making; as such, the test does not return 'the one and only answer' to a research question. It is recommended that mechanised interpretations of the results, such as relying uncritically on the 5% threshold rule to draw conclusions, be avoided.**
- **A small ('significant') *P*-value means only that it is unlikely that the observed difference is due to pure chance. It does not mean that the difference is clinically relevant (use confidence intervals to assess this), and neither does it imply that there is a causal relationship between the two phenomena (it may still be possible that the observed difference is due to the presence of confounding factors or more generally to some form of bias).**
- **In particular, consider that significance increases with sample size. Even a very small difference at the population level ('in nature') will be highly significant in a very large sample. Additionally, only a very large difference at the population level will be significant in a small sample.**
- **A nonsignificant *P*-value does not prove that there is no difference; it only indicates that the data do not provide sufficient evidence to reject the hypothesis that there is no difference. Use tests for equivalence or non-inferiority to support a hypothesis of no difference.**
- **Finally, be aware of the problem of multiple testing. The more tests you apply in an analysis, the higher the probability of 'false discovery', that is, the higher the risk that just by pure chance, the data will show a 'significant' difference in the population when there actually is none.**

2. Planning a study in SCT

2.1. Endpoints: definition

As briefly introduced in Section 1.1, the main objects of interest in SCT are the occurrence in time of certain events that are either certain to occur, such as death, or possibly prevented by the occurrence of competing events, such as relapse, which has death without prior relapse as a competing risk. Individual survival times and times-to-events with competing risks may not yet be observable at the last follow-up, leading to censored observations. Defining an endpoint of this type implies fixing when to start and when to stop the clock that is measuring the time-to-event of interest. This section will provide more methodological insight into these issues.

We will make examples out of an SCT context for malignant diseases, thus introducing commonly aligned endpoints (OS, RFS, NRM (non-relapse mortality) and so on). However, we would like to stress that specific situations may require definitions, or case-specific endpoints, that are different than the ones that we will provide.

2.1.1. Survival times and censoring

Defining OS is relatively simple: it is the duration of survival from a certain point in time until death. In these sections, we will for the moment skip the issue of fixing the starting point, and consider the (first) transplantation as the time of origin as is normal in most SCT studies. The final event is unambiguous and certain to occur; the only uncertainty is when.

Nonetheless, in any study, the duration of the follow-up period is not infinite, and thus, it is possible that some patient was alive at the last follow-up and the actual total survival time is not observable. Another reason for not observing the actual survival time is loss to follow-up; some patients, for example, stop having hospital check-ups or move to other institutions, so they are no longer observed after a certain time point despite the fact that, in principle, the observation should be on-going. However, it is important to remember something that may appear trivial in this context but which may be less obvious for other investigations: All cases with insufficient or incomplete follow-up will experience death—the event of interest—at some unknown time after the last follow-up. In other words, at the moment of last follow-up, they are still at risk of failure. Although we ignore the actual duration of their survival, we know it will be longer than the duration of their follow-up. For this reason, their survival time is not 'missing', but it is 'censored'.[13] The branch of statistics called survival analysis proposes methods that correctly include this incomplete information in the analysis.

The majority of statistical methods for survival data, and in particular the basic methods illustrated in this document, rely on the assumption that censoring is 'independent and not informative' with respect to the outcome of interest,[14] which, under the perspective of interpretation, means that being censored at a certain time should not depend on the risk of experiencing the event(s) at that time or later. This means that those cases that are still under observation and at risk of the event(s) of interest at a certain time are representative of the cases with the same characteristics that are censored at the same time, at least insofar the current and future risks are concerned. Violation of this hypothesis leads to biased results. For example, applying the Kaplan–Meier estimator (Section 2.2.1) to censored cases with a higher risk than the uncensored cases leads, as one might intuitively expect, to an overestimation of the survival probability.

[13] What is described in this section is actually right-censoring, observations that are incomplete in the sense that, if we indicate with T the time to the event of interest, and with C the time to loss to follow-up, we have $T > C$. Situations with left- or interval-censoring are less common in practice, and the methods for addressing them become more complicated. Refer to Kalbfleisch and Prentice[43] (Section 3.2) for the theory on different patterns of censoring.

[14] To be precise, this assumption can be slightly relaxed (see Kalbfleisch and Prentice[43]), but the substantial, interpretative meaning is the one reported here.

Common mistakes made with censoring

The nature of censoring implies that censoring observations at the occurrence of some ‘nuisance’ event different from a true loss to follow-up must be avoided because those events usually correspond to a change in the risk of the patient. The following examples of these common mistakes introduce issues that will be discussed in the later sections:

- Censoring cases that are dead without relapse to estimate the probability of relapsing: this is a case of competing risks, and must be treated with specific methods (see Sections 2.1.2 and 2.2.2).
- Similarly, censoring a cases case when the cause of death was different than the specific cause of interest, such as censoring a non-disease-related death when the aim is estimating the incidence of disease-related mortality (again, this is a competing risks problem).
- In a PCT on survival, censoring patients when they go ‘off-treatment’ for toxicity: these patients can be expected to be at higher risk of death than those who do not stop treatment for toxicity. See the discussion on the intention-to-treat (ITT) principle (Section 2.3).
- Censoring patients when they have a second transplant on the basis that this event modifies the course of the disease, while you are only interested in survival after one transplant: usually, second transplants are given for events such as relapse and graft failure, which identify patients as being at higher risk for death than those not requiring a second transplant. The management of second transplants or other ‘intermediate events’ will be discussed in Section 2.1.6.

A survival-like endpoint is defined as the time from an origin to an event that is certain to occur. An observation is censored when the event of interest was not observed during the follow-up period; however, the patient is still at risk of the event, which will occur at some unknown time after the last follow-up.

When defining an endpoint and what will constitute a censored observation, ask yourself: ‘Does the fact that the patient is censored at a certain time indicate that his/her risk is higher (or lower) than the risk of an identical patient which is still on follow-up at that time?’ If the answer is ‘no’, then your definition of the censored observation is correct.

More could be said about censoring, particularly as there are several potential problems related to the manner in which follow-up is conducted in real studies. For insights, see Marubini and Valsecchi,[2] Section 3.5.3.

2.1.2. Competing risks

In Section 1.1, we saw an example of a situation characterised by competing risks with reference to the analysis of relapse; this example is useful to introduce more remarks and further issues.

Example 1: Relapse, non-relapse mortality and RFS.

Patients with acute leukaemia are transplanted when in CR and may relapse afterwards; patients may also die, for whatever cause, without ever experiencing relapse. This latter situation is referred to as non-relapse mortality, NRM.[15] NRM is thus a competing event of relapse. Relapse and NRM can be seen as two different causes of ‘failure’; one of them will occur, but only one, and could be observed during a sufficiently long follow-up (Figure 1).

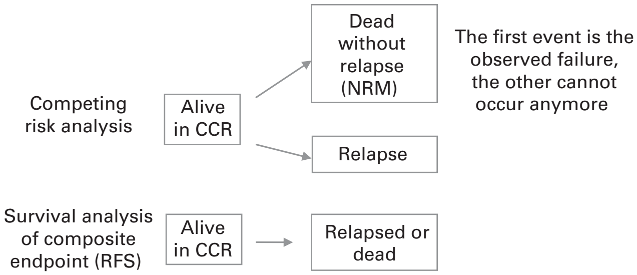
Competing risk analysis
Alive in CCR
Dead without relapse (NRM)
Relapse
The first event is the observed failure, the other cannot occur anymore
Survival analysis of composite endpoint (RFS)
Alive in CCR
Relapsed or dead

Figure 1 Definition of endpoints.

Obviously, patients who die without prior relapse will certainly not experience relapse afterwards: the risk of relapse after NRM is zero. Equally, although it may seem counter-intuitive, patients who relapse will certainly not experience mortality without relapse afterwards;[16] these patients’ risk of NRM is zero after relapse. This highlights the fact that the occurrence of a competing risk cannot be considered as censoring for the time-to-event of interest. In fact, in any time-to-event context, censored patients are by definition only those still at risk of failure, while when one competing event does occur, patients are no longer at risk for the other event.

In the example of relapse and NRM, the censored observations are those of the patients who at the last follow-up are alive and have never experienced relapse before; they are still at risk for both events, one of which will occur with certainty after the last follow-up at some unknown time.

Competing risks exist whenever an event of interest can be precluded from occurring, even with an ‘infinite’ follow-up period, by another event(s). Treating the time-to-event as a survival time and censoring at the occurrence of a competing event is in general improper, and usually leads to biased results.

Several events of interest in SCT research have competing risks. Death may prevent observing engraftment, GVHD and CR achievement (‘death’, in this case, must be intended as ‘death with no prior event’).[17] Depending on the clinical context (type of disease and so on), there could be additional competing events, such as second transplant and perhaps chimerism for engraftment, relapse for GVHD, progression for CR achievement and others.

[15] In this context, NRM is more appropriate than TRM, which stands for transplant-related mortality or, in non-transplant settings, treatment-related mortality, because TRM suggests classifying the cause of death as either transplant-related or not, while here we specifically mean ‘death without ever experiencing a prior relapse’. See also the remarks below in the ‘common mistakes’ subsection.

[16] In fact, by NRM we mean mortality without prior relapse, and not mortality due to causes other than relapse.

[17] Although it is customary to report these events in terms of simple percentages of occurrence and to neglect the competing events, this is often the wrong approach. Appropriate approaches are discussed in Section 2.2.3.

Whether an event acts like a competing risk must be discussed between the statistician and the clinical investigators of the study.

Another context involving competing risks is the analysis of the causes of death. This type of study is always affected by the chance that the quality of the information on the exact cause of death may be insufficient. In some cases, there is an objective difficulty in attributing the cause of death. For example, suicide or car accidents could be classified as belonging to a generic 'other cause' category or they could actually be 'treatment-related' causes if they are attributable to some sort of neuropathy. Thus, it is advisable to consider a few, relevant and plausible causes of death for this type of study.

Common mistakes made with competing risks

In the literature of some medical fields, the presence of competing risks is ignored in the definition of the endpoints and/or during statistical analysis. In SCT literature, this is infrequent, but confusion sometimes occurs in the definitions of competing events, or the authors neglect to describe precisely in the publication which events are competing and which are the censored observations. For example, it is possible to find in the literature the same term, transplant-related mortality (TRM), used for two different endpoints, one defined as the NRM in our Example 1 and competing with relapse occurrence, and the other defined as death attributable to transplant-related causes and competing with deaths that are non-transplant-related (in studies on causes of death).

2.1.3. Composite survival endpoints

Example 1 (continued).

Without distinguishing the cause of failure, relapse or NRM, in this situation it is also of interest to study the time to relapse or death, whichever comes first. The combined failure event will be observed with certainty after a sufficiently long follow-up. Thus, the defined endpoint is a survival-like outcome, usually called 'relapse-free survival' (RFS; Figure 1). This endpoint is meaningful as it computes the time spent in continuous CR after transplantation. For RFS, patients alive at last follow-up who never experienced relapse are the censored observations.

RFS is an example of a 'failure-free' survival-like endpoint. It measures the time from a certain origin to a composite event, the first of two or more events that usually represents a failure of the therapy, such that at least one of the events occurs with certainty (notice that when death is a component of the combined event this requirement is always satisfied).

Other examples can arise in situations where it may be of interest to stop the clock measuring the 'failure-free' survival time, such as the development of GVHD, fungal infection, graft failure, no response and so on. In PCT, it may be necessary to adopt particular definitions for the endpoints to manage events such as the patient's failure to be compliant with the assigned treatments, violations and drop-offs (see the issue of ITT in Section 2.3). These type of endpoints may be generically referred to as 'event-free survival' (EFS), or they may receive specific names ('fungal-free survival' and so on). Terminology may in fact generate confusion if inappropriate or imprecisely described, as illustrated in the next subsection.

Survival-like endpoints can be defined by means of a composite failure corresponding to the first of a number of events such that the failure occurs with certainty in a sufficiently long follow-up time. They are thus analysed similarly to OS.

RFS, DFS, PFS … Potential semantic pitfalls

Very often the expressions 'relapse-free', 'disease-free' or 'leukaemia-free' and 'progression-free' are used equivalently when considering an outcome such as RFS (defined in Example 1). However, to avoid confusion, it is necessary to apply a definition and then use the appropriate terminology that corresponds exactly to the clinical situation being analysed, especially as the results of different studies are often quickly read and compared without carefully reading the Materials and Methods sections of the papers.

The term disease-free survival (DFS) is rather common and is used as an alternative to RFS; thus, it is used for patients in CR at the start time and with relapse or death as the final time. In semantic terms, an obvious restriction is that the word 'disease' refers to the specific disease under investigation (patients may suffer from several diseases during their lives). The term leukaemia-free survival (LFS) is used in place of DFS in the studies on leukaemia. Additionally, it should be specified that DFS extends until the time of the first recurrence and does not represent the total time without disease (see the current LFS, Section 2.1.7).

The term progression-free survival (PFS) is proper for cases where patients are not in CR at the start time, making progression a failure of interest along with death; relapse is included, being a subtype of 'progression' for patients who start or reach CR.[18] This type of situation occurs in SCT in diseases such as lymphoma or multiple myeloma where patients can be transplanted with active disease, or in autoimmune diseases. Notice that in these cases RFS could also be analysed by restricting analysis to patients who were in CR at transplantation or who achieved CR later by taking into account the delayed entry (Section 2.1.4).

Common mistakes made with composite endpoints

- When defining an endpoint based on a composite failure, it is fundamental that the definition is homogeneous and applicable to all patients included in the analysis. Below are two examples of incorrect definitions:
    - ○ Consider the analysis of a group of patients with either autologous or allogeneic transplantation where GVHD is considered a failure in the EFS along with relapse and death. This is questionable because GVHD can only be experienced by the allogeneic patients.

[18] There is a semantic problem also for the event competing with relapse/progression: usually the acronym NRM is used, but more precisely we should use something like NRPM. In any case, it is important to specify (in presentations and in Statistical Methods sections) that we mean death without prior progression or relapse.

  - Again using autologous and allogeneic patients, consider the case where a second transplantation is neglected (or censored) for the autologous group and considered a failure for the allogeneic group. Unless specific reasons justify this approach, this should not be done; the same type of event should not be coded differently in different cohorts of patients.
- Non-homogeneity should also be avoided with respect to the different timing of events.
  - Consider the case where an EFS endpoint is defined to include no response, relapse (if the patient responded) and death as failures. However, there are two treatment groups, and due to different durations of therapy, the achievement of a response is assessed later in group A than in group B. The comparison in terms of EFS is biased, with group B being disadvantaged by construction.

Competing risks or composite survival endpoints?

In the presence of competing events, the investigator can usually choose whether to perform a competing risks analysis, use a composite survival-like endpoint or perform both analyses.

However, there are exceptions where the use of composite events are not applicable. For example, when the event of interest is favourable for the prognosis (engraftment, achievement of CR), because death is always a competing event, the combined endpoint (for example, time to engraftment or death) is not really meaningful. An EFS endpoint could be defined with ‘failure to achieve engraftment within time x’ as a component of the event.

When both analyses are meaningful, it is a matter of relevance and interest, which one should be the main target, although it is always advisable to use both. Analysing only the combined endpoint lacks potential for understanding the phenomena, but only performing the competing risks analysis misses the necessary synthesis. It is particularly important to avoid analysing only one of two competing risks. In Example 1, although the major interest may be the relapse rate, NRM should not be neglected; it could be a good choice to present both the cumulative incidence curve of relapse (estimating the relapse rate per time, Section 2.2.2) and the RFS curve that includes NRM. Stacked curves (Figure 2) provide an excellent synthesis of the entire situation.

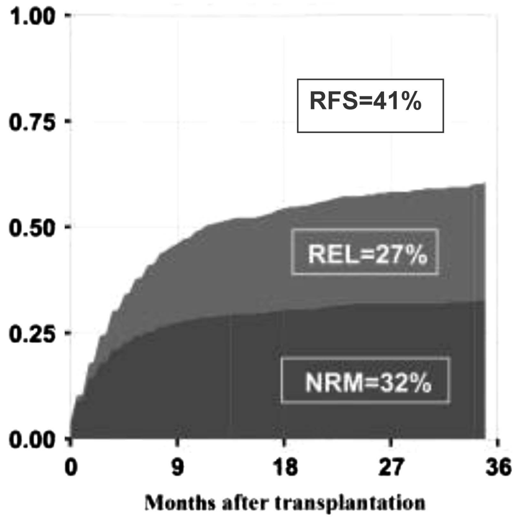

Figure 2 Stacked curves for competing risks setting (relapse and non-relapse mortality).

The choice is particularly relevant in clinical trials where the number of patients to include is usually based on the target of evidence required for only one (primary) endpoint. In the case of a comparison of treatments aimed at reducing the rate of relapse, choosing the latter as the primary endpoint may be natural, but it is questionable because—and this is making an extreme example to highlight the concept—strictly speaking, the best way to obtain a very low relapse rate would be to have an extremely high death rate after transplantation. The problem is realistic, however, when highly toxic treatments are involved. Of course, we do not mean to imply that clinical investigators would neglect the risk of mortality, but we stress the need of taking NRM explicitly into account during the planning phase. At the very least, the protocol should make clear statements regarding the expected non-relapse mortality and/or RFS, and then, the analysis should check for it. Actually, because it is difficult to design a trial controlling for two different endpoints, using RFS directly as a primary endpoint to fix the sample size may be preferred, but there is a relevant drawback: Because of the trade-off between toxicity and efficacy, the two treatments may produce a small difference in terms of RFS, where one has lower treatment-related death but also less efficacy in preventing relapse, and the second displays the opposite. Thus, the expected difference could be small, and a very large number of patients may be required to detect it. If the two treatments have similar mortality instead, then RFS appears to be the best candidate for a primary endpoint.

2.1.4. When to start the clock*

In Sections 2.1.1 to 2.1.3, we discussed issues that are related to ‘when to stop the clock’ for studies involving time to an event of interest. The clock stops at the occurrence of the event of interest, or of the competing events, or at the date of last follow-up if no event occurred. But when does the clock begin?

There is usually a clearly identified situation when the clinical history of interest begins. For example, in SCT studies it is the transplant, and in solid oncology it is the diagnosis or the start of therapy. In a randomised clinical trial, it should be the date of randomisation. A general definition for the start time is the *first occasion when the patient is at risk* for the event of interest in the context of the study, and we can consider this to be an ‘entry time’ into this status (‘being at risk’). Because it is natural to compute the time-to-event as the time elapsed since this starting situation, we tend to identify the entry time as the origin (the time zero) of the time scale. However, this is not always the case. To illustrate this issue, as well as the relevance of choosing an appropriate method of analysis, we must consider the concepts of time scales and delayed entry.

A time scale is a time axis along which the risk of failure varies. All the methods of time-to-event analysis illustrated

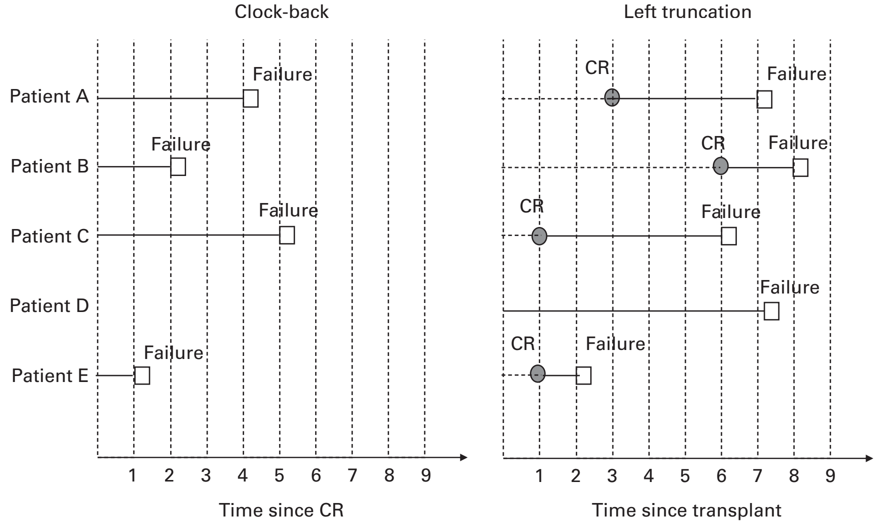

Figure 3 Time scale issues: left-truncation and the 'clock-back' approach.

in this document use only one relevant time scale for evaluating the risk of failure: assuming that two 'identical' patients assessed at the same time along a given scale necessarily have the same risk, while the risk is (potentially) different for another identical patient who is assessed at another time.

The choice of the time scale is crucial for analyses because the statistical procedures are based on ranking the observed failure and censoring times along the reference time scale and subsequently identifying the risk sets, the set of individuals at risk at each point in time (the concept is better illustrated in the example below). The differences in terms of relative risk at time t that are attributable to patient characteristics are evaluated among the individuals belonging to the risk set at that time. If it is possible to choose between different time scales, this choice affects the risk sets, thus having an impact on the results of the analysis.

We will illustrate with an example situation in which there is a problem with the choice of the reference time scale and how this problem can be approached.

Example 2: Time scales in the analysis of RFS after achievement of CR.

In a transplantation setting, the reference time scale is usually the time since transplant. However, in our example study, we are interested in RFS and CR is achieved after transplant, thus we have to consider that the patient starts being at risk only when CR is achieved. In other words, the patient enters the status 'at risk' at a time E, which is different from patient to patient and later than the time of transplant. Obviously, it is not sufficient to exclude from the analysis those patients who did not achieve CR. All respondent patients should be excluded until they reach CR, because by definition they could not experience the risk of relapsing before CR. This problem can be solved in two ways.

One approach is to re-compute the time-to-failure since the moment CR was achieved. This is known as the clock-back approach because the 'clock' measuring survival is set back to zero for all patients at the time they achieved CR status. Two identical patients thus have the same risk when evaluated at the same time since achieving CR, even if one was transplanted 1 month before and the other 2 years before.[19]

The second approach[20] is to use the time since transplant as the reference time scale, but acknowledge the delayed entry in the status of being at risk of failure, which is achieved by modifying the risk sets as shown below. In terms of probability distributions, this corresponds to imposing the condition that the failure time T has to be larger than the entry time E (in this case, time to CR), which is called left-truncation.[21]

We can demonstrate the two approaches with an example using the data for five patients, as illustrated in Figure 3. In the case of the clock-back approach, the order of failure times from the smallest to the largest is t_{E}, t_{B}, t_{A} and t_{C} (t_{D} is not defined), and the risk set at time t, $R(t)$, is the set of individuals whose observed failure time is larger than or equal to t. Thus: $R(t_{\mathrm{E}})=\{\mathrm{A, B, C, E}\}$, $R(t_{\mathrm{B}})=\{\mathrm{A, B, C}\}$, $R(t_{\mathrm{A}})=\{\mathrm{A, C}\}$ and $R(t_{\mathrm{C}})=\{\mathrm{C}\}$. Notice that the dimension of the risk sets decreases in time, with all patients belonging to the first risk set, and subsequently leaving the sets for failure (or censoring, although all observations are complete in this example).

In the case of the left-truncation approach, the order of failure times from the smallest to the largest is t_{E}, t_{C}, t_{A} (t_{D}) and t_{B}; and the risk set at time t is the set of individuals whose entry time E is lower than t and whose observed failure time is larger than or equal to t. Thus, we have: $R(t_{\mathrm{E}})=\{\mathrm{C, E}\}$, $R(t_{\mathrm{C}})=\{\mathrm{A, B, C}\}$, $R(t_{\mathrm{A}})=\{\mathrm{A, B}\}$ and

[19] This is true unless the time between transplant and CR is included among the covariates.

[20] This approach is also sometimes termed clock-forward, especially in the framework of multi-state models[5].

[21] In practice, not all statistical software programs allow the researcher to apply left-truncation. With this approach, the survival data need to be arranged in the counting process form, that is, they must be represented as a triplet (entry time, final time and failure indicator).

$R(t_B) = \{B\}$. Notice that patients A and B did not belong to $R(t_E)$ because they had yet to enter the condition of being at risk, at time t_C the risk set was larger than at time t_E, and patient D does not enter any risk set because his entry time E is not observed (he does not achieve CR).

If we erroneously kept the time since transplantation as the time scale without correcting the risk sets for delayed entry, the first of them would have been $R(t_E) = \{A, B, C, (D), E\}$; patients A and B (and D) would thus appear to be at risk of failure during a period in which they were not (before achieving CR), which clearly can lead to the wrong assessment of relative risk.

Delayed entry can occur every time you select patients on the basis of information that is known at a time later than the natural origin, or in more general terms, when the condition of being at risk arises during the follow-up started at the natural origin. This requires setting the clock back or applying methods for delayed entry (left-truncation).

Of course, if the time to entry in the risk status is the same for all patients (for example, being at risk of chronic GVHD starts at day 100 from transplant when the traditional definition is applied), then the two approaches return the same risk sets.

Notice that all transplant histories are actually disease histories started before transplant, at diagnosis: we usually apply the clock-back approach and include time from diagnosis to transplant as a covariate in the analysis.

2.1.5. Creation of outcome variables in practice

Once an endpoint of interest is chosen, new columns must be created in the database that will specify the outcome variable in the statistical procedures you will apply. We present an example, although depending on the method and on the software used, the inputs or coding required may be different.

For survival and competing risk situations, we need two columns, one for the time and another for the status (we will in fact disregard the issue of left-truncation, which would require an extra-column for the entry time):

- When the event of interest occurred, the time variable is equal to the time interval between the origin and the event, and the status variable is 1; for endpoints with composite failures, the time variable is computed from the origin to the first failure that occurred.
- If there are competing risks and a competing event occurred, the time variable is equal to the time interval between the origin and the competing event that occurred, and the status is equal to 2 / 3 / … (one code number, 2 to k, where you are interested in k competing risks).
- In both situations, if no event occurred, the time variable is equal to the time interval between the origin and the last follow-up, and the status variable is equal to 0 (censored cases).

Example 3: Computing the outcome variables from the data

The following schema represents the course of the disease for seven patients in a study with a maximum duration of 18 months. Time is measured in months, the origin is transplantation, R indicates relapse (all patients were transplanted while in remission), D indicates death, and A indicates the date the patient was last seen alive (Figure 4).

The data are:

Patient id	Relapse occurred (no = 0, yes = 1)	Time to relapse	Death occurred (vital status; no = alive = 0, yes = dead = 1)	Time to death or last contact
1	0	–	1	14
2	0	–	0	18
3	1	8	1	18
4	1	8	0	18
5	0	–	0	10
6	1	4	1	16
7	1	14	0	17

The last two columns represent the status indicator (1 = event occurred, 0 = censored) and time columns for OS, respectively. In addition, in the following table we compute the columns necessary to define the outcome for the RFS, the analysis of relapse incidence and non-relapse mortality in a competing risks setting, and survival after

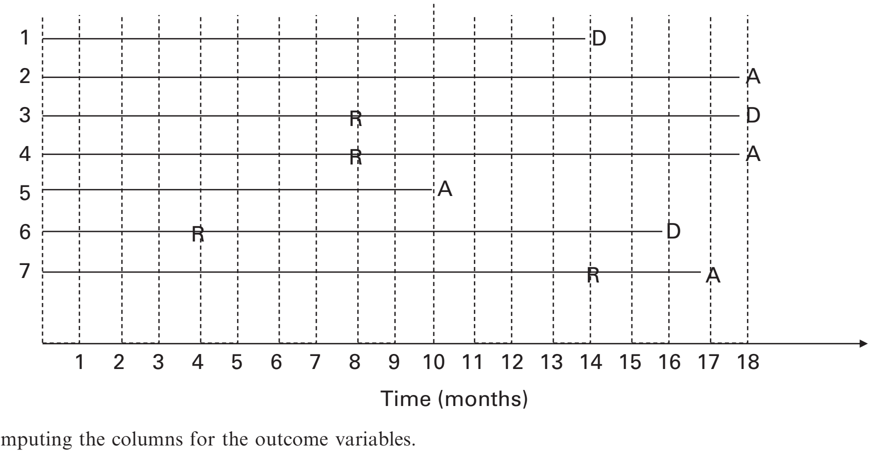

Figure 4 Example 3, computing the columns for the outcome variables.

relapse. The latter is an example of an outcome with a different time origin, relapse, so the clock is set back to 0 at the time of relapse. Patients without relapse are excluded from the analysis.

Patient id	Status indicator for relapse incidence/ NRM	Status indicator for RFS	Time variable for RFS or relapse incidence/ NRM	Status indicator for survival after relapse	Time variable for survival after relapse
1	2	1	14	–	–
2	0	0	18	–	–
3	1	1	8	1	10
4	1	1	8	0	10
5	0	0	10	–	–
6	1	1	4	1	12
7	1	1	14	0	3

2.1.6. Second transplant and intermediate events in general*

Very often investigators interested in the occurrence of a certain final event struggle with the statistical management of other events that occur during the course of the disease.

One situation is when these other events are considered 'nuisance' events in the sense that the investigator is not interested in them, neither as outcomes nor as prognostic factors, but feels that they affect the object of interest in some way, and thus wants to disregard their occurrence and the subsequent disease history. A typical example is the analysis of the long-term effects of chemotherapy for acute leukaemia patients at onset when some of them received allogeneic transplants at some point. A second case is when the event actually represents a good or bad outcome of the disease process, and thus, there is an interest in considering them as competing risks or as components of some composite EFS, as in many of the examples already seen. A third case is when the investigator is interested in understanding whether the disease history is really affected by the occurrence of an intermediate event, and if so, how. A typical problem is investigating the role of second transplants, or the possible change in prognosis when events such as GVHD or response occur.

Censoring is (usually) a mistake

When an investigator wants to analyse a certain life history, but feels that the intermediate event causes modifications with respect to what would have been observed without its occurrence, the tempting 'solution' is to censor the observation at the time the 'nuisance' event occurred. As was described in Section 2.1.6, the analysis is unbiased only when the patients remaining at risk are fully representative in terms of the subsequent risk of failure of those being censored at the same time. This is hardly the case when there is censoring at the occurrence of events other than the event of interest. Consider the following two examples in the context of oncohematology and SCT:

- In the example of the analysis of the long-term effects on OS of chemotherapy for acute leukaemia patients at onset, the investigator applies a censoring at the administration of allogeneic transplantation with the argument that due to GVHD, chimerism, and so on, the transplanted patients are not homogeneous to the rest after transplantation. Because (usually) transplantation is given only to patients in CR after induction, as censored cases they have at that time a lower risk of death than the remaining population, which includes patients who are non-responsive at the same time. As a result of this 'informative' censoring, the OS probability is underestimated.
- In a registry-based SCT study with OS as the main endpoint, the investigator applies a censoring at the occurrence of second transplant with the argument that the interest is only in the outcome of the first transplant. In the likely case that second transplants are given as a consequence of an event that has a negative impact on survival, such as relapse or graft failure, then at the time of second SCT patients have a higher risk of death than those who do not require a second transplant at the same time. As a result of this 'informative' censoring, the OS probability is overestimated.

Generally speaking, the censoring of patients at the time of second transplant while analysing OS may violate the requirements of proper censoring. In fact, in the majority of cases the administration of second transplant is associated with a particular risk status (high risk of death, if given as salvage treatment, or low risk of death, when given in CR), and the censored cases would therefore not be homogeneous with the non-censored cases. It may be an independent censoring, for example, in the case of a clinical trial where the second SCT is given according to randomisation and independently of the current response/risk status of the patients.

What to do instead of censoring?

In the example of the analysis of long-term OS from induction, it is perfectly appropriate to completely neglect whether the patient received an allogeneic transplant or not. The fact that the patient survived long enough to achieve CR and then an allogeneic transplant, thus earning an additional short-term risk followed perhaps by a reduction of risk, is part of the treatment history after the initial induction therapy and contributes to the final outcome. This reasoning is substantially based on the ITT principle (Section 2.3).

In the example of second transplants given following relapse or graft failure, the appropriate perspective may be to consider second transplantation as an outcome of interest in itself, representing a type of failure or a surrogate for failure. In this case, a second SCT can be managed as a competing risk or as part of a composite failure (Section 2.1.3). The same approach can be used in defining the endpoint in a clinical trial that prescribes starting a therapy to prevent relapse according to the levels of certain biological markers; the initiation of such therapy may be included among the events indicating failure. Regarding second transplantation, notice also that it appears appropriate to treat it as a failure in all analyses where the unit investigated is the "graft" and not the patient. This is the usual approach in studies on organ transplant or heart valve implants where the interest is in analysing the duration of survival with one organ or the duration of the device, respectively, stopping the clock if and when the patient receives another organ or device (the ‘life’ of the organ or device ends). The analysis of engraftment in SCT actually follows this logic.

Analysing the ‘effect’ of an event

So far in this section, we have seen the intermediate event as something that in some way terminates a disease history. However, the changes in the status of the patient that are associated to (or possibly caused by) the intermediate event are often objects of interest as prognostic factors. Examples include the investigation of the benefit of second transplantation or the change in prognosis when the patient achieves CR or develops GVHD.

A very typical and serious mistake[3,4] is comparing the outcomes between those patients who experienced the intermediate event and those patients who did not, without taking into account the fact that the intermediate event took place during the follow-up and after the start of the clock. For example, an investigator compares the curves of the OS probability from the first transplants of two subgroups, patients who later received a second transplant and patients who did not. This mistake involves a classification of patients based on the future. All of the patients who died early, and because of that, could not receive a second transplant automatically belong to the ‘no second SCT’ group, which is thus disadvantaged by construction. This is an example of time bias. In these situations, the statistical analysis must in some way adjust for the fact that to experience an intermediate event, the patient first has to survive during a ‘waiting time’. The principle is that to perform an unbiased comparison between patients with and without the event, it must be done at equal ‘waiting time’.

One approach is to make a landmark analysis, selecting only the patients still at risk at time x and then defining and comparing two groups, those who received a second transplant before time x, and those who did not. This latter group includes both patients who received a second SCT after time x and those who received only one transplant.[22] Once the two groups are set, there must be a proper accounting for the delayed entry either by setting the clock back or using left-truncation (Section 2.1.4; because the entry time is the same for each patient, the two approaches are equivalent). Here, the classification is not based on the future but on the past (until time x). The advantage of this approach is that the analysis is quite simple, and if choosing the clock-back approach it can be performed using ‘standard’ methods such as Kaplan–Meier curves and the log-rank test. The disadvantage is that it is rather restrictive with respect to the aims of the investigation, the comparison may be influenced by the choice of the time threshold x and the analysis does not consider the first part of the story of the disease, that is, early failures.

Another approach is to include the occurrence of the intermediate event as a time-dependent covariate in the Cox regression (Section 4.3.3). A time-dependent covariate is a variable that changes its value over time. For example, the variable that indicates the occurrence of a second transplant changes from the value 0 to the value 1 at the time the second transplant is given, and it is always equal to zero if a second SCT is not performed. The model allows the investigator to compare at each point in time those patients who at that time had a second SCT to those who did not. This is a more general analysis than with the landmark approach, and all cases are included.

The limitation of this approach is seen when we study the effect of several prognostic variables including the occurrence of the intermediate event. Some (time-fixed) characteristic at transplant can affect the chance of experiencing the intermediate event (time-varying variable), and the Cox model including both variables fails to detect the global impact of the characteristic because part of its effect is represented by the occurrence (or not) of the intermediate event. For example, treatment A produces a higher risk of death than treatment B but also has a greater chance of having a second transplant than treatment B, whereas the latter reduces the mortality. Globally, the effect of treatment A could be comparable or even superior to treatment B, but the Cox model returns only an estimated effect of A assuming that the second transplant was given (or not given). This limitation is overcome by using multi-state models.[5,6] With this approach, each part of a disease history with one or more intermediate events is modelled using a Cox regression (or indeed, any other estimation technique), and the results are combined to estimate the probability of each possible outcome in time. For an example of the application of both time-dependent covariates and multi-state modelling in an investigation on the role of second SCT, see Iacobelli *et al*.[7] Multi-state models have high potential, but the application is rather complicated; we therefore suggest that this approach be left to statisticians who are experienced in the field. EBMT guidelines dedicated to multi-state models[8] and an extensive literature are available. It is also worth mentioning an alternative based on dynamic prediction.[9]

Aside from which statistical method is used, it is important to comment on the interpretation of analyses of the ‘effects’ of intermediate events. Because these events are themselves outcomes of the disease process (unless, for example, they correspond to the administration of treatments given according to a protocol and independent of the status of the patient at that time), it is hazardous to speak of causal effects (see also ITT analyses, Section 2.3). For example, the sentence ‘Second transplant reduced the risk of death’ suggests the presence of a causal effect, while the observed benefit could be explained by some type of auto-selection process such that the ‘fittest’ patients, those having a lower risk of death, also had higher chances of receiving a second SCT. A more appropriate sentence would therefore be a ‘descriptive’ one, such as ‘Patients who received a second transplant showed a lower risk of death’.

When a ‘secondary’ event occurs during the course of a disease, there are very few situations where treating it as

[22] The time threshold *x* should be chosen so that the most relevant part of the information is acquired by time *x*.

censoring is either meaningful or appropriate. Instead, evaluate whether you should ignore the secondary event under an ITT perspective; treat the secondary event as a failure in a competing risks setting or as a component of a combined endpoint; or investigate its role or adjust for its occurrence using landmark analysis, Cox regression with time-dependent covariates or multi-state models. In the last case, be sure to avoid misinterpreting the relationship found between the intermediate and final event, unless causality is truly established.

2.1.7. Other issues regarding relapse and competing risks in general*

'I am only interested in relapse if mortality did not exist': Latent failure times

The analysis of relapse is a typical case in which the investigator is very much interested in that specific event and not at all interested in the competing one (mortality without prior relapse, NRM). This happens in all clinical contexts where mortality is negligible or otherwise 'adequately controlled', at least in the first few years. In simplified terms, we could say that the investigator would in some way like to have evidence relating only to relapse, as if NRM could be eliminated entirely. Early statistical approaches to event-history data in the presence of competing risks were in fact based on what was called the latent failure time approach. In this approach, it was presumed that there was a 'hidden' survival process for each possible cause of failure, all processes were independent (this was a necessary condition as otherwise the model was not applicable in practice) and only the first failure occurring among all competing events could be observed. In other words, even after NRM was observed, the 'clock' measuring the time to relapse was continuously running, only it was no longer visible and the final time to relapse remained unknown. According to this approach, the real object of interest in the analysis of relapse is the latent survival-like endpoint or, in formal terms, the corresponding marginal survival function.[23] Apart from practical limitations, this approach is misleading because a competing event cannot be eliminated from reality.

'Relapse is not a permanent status': Current RFS

While in competing risks analysis the disease history ends with the occurrence of relapse and endpoints such as RFS are used to compute the time until first relapse, in certain frameworks, such as in SCT for chronic myeloid leukaemia where patients who relapse may achieve CR again, then have another relapse and so on, it is interesting to evaluate the probability of being alive without relapse after relapse occurs for the first time. In the context of leukaemia, this type of endpoint is known as current LFS,[10–12] and the object of interest is the probability of being alive and leukaemia-free at any point in time, regardless of whether the patient is in first, second or subsequent post transplant remission. This is an important parameter for evaluating therapeutic strategies that incorporate planned post transplant treatment of relapse such as DLI. It can be estimated within the multi-state approach (Section 2.1.6) by building models for all transitions from CR status to relapse, then to CR again and so on.

2.2. Endpoints: statistical methods

Depending on the type of endpoint and the objectives of the study, a large variety of statistical methods are available for analytical purposes, and new methods and models appear almost daily in the statistical literature. This section briefly introduces only the most common methods used and, in particular, those methods that are currently used by EBMT statisticians or which are commonly found in the SCT literature. Specific problems or clinical questions may require more advanced methods than the 'standard' methods presented here. Apart from a few 'technical' aspects, this section addresses the assumptions, potentials and limits of each method whose understanding is fundamental for a critical appraisal of the results of an analysis.

2.2.1. Survival-like endpoints

When the endpoint of interest is a survival time, such as OS and RFS, the relevant objects that describe this endpoint are the survival function $S(t)$ and the hazard function $h(t)$ and their derived quantities. At this point, we will not illustrate them as mathematical objects with specific formal properties but will instead focus mainly on their clinical meaning. For the discussion, we will use the generic words 'survival' and 'death' (the latter is used instead of failure), but obviously, what is said here can be generalised to all failure-free survival endpoints.

Survival function

The survival function $S(t)$ states for each time t the probability of surviving beyond time t, or the percentage of patients who are expected to still be alive at time t. At the time origin, $S(0) = 1 = 100\%$; for increasing t, $S(t)$ decreases, and with a long enough follow-up, or theoretically at infinite time, $S(t)$ decreases to 0, as death is an event that occurs with certainty.

The complement to 1, $1-S(t)$, which is sometimes called the one-minus-survival function (OMS), is simply the probability of dying before time t, or the cumulative mortality rate[24] at time t.

The survival function is estimated by the Kaplan–Meier or 'product-limit' method. For an example of estimated survival curves, see Section 3.5.2. Another method that provides somewhat different estimates is the 'actuarial' or 'life table' method, which could be used in the case of grouped ('discrete time') survival data.[25]

[23] Recent statistical methods approach the estimation of the marginal survival function. Such efforts can be found in the work of Fine and others,[44] who introduced the term 'semi-competing risks' for settings where the role of two events is asymmetric, as it is here where death prevents relapse ('truly competing' events), but the converse is not true.

[24] In statistics and epidemiology, 'rate' refers to a quantity that is not interpretable as a probability. However, in these Guidelines we follow the traditional practice of the medical literature and use this term to in fact mean 'probability'.

[25] An interesting example of grouped survival data is when it is not known precisely when the event occurred, but only that it occurred during a certain period, such as between two different scheduled follow-up times. Useful remarks can be found in the books by Kalbfleish and Prentice[43] and Marubini and Valsecchi.[2]

The time t^* where the estimated survival curve is equal to 0.5 (50%) is the median. The interpretation is that half of the patients die within the median time t^*, or that half of the patients survive at least beyond t^*. The majority of statistical software programs provide the estimated median with a 95% confidence interval. A more complete synthesis of the survival curve is a series of estimated survival probabilities at specific points in time with their confidence intervals.[26] It is also very important to look at the number of patients still at risk at each time. At the end of follow-up (that is, at the tail of the curve), there are usually very few patients still at risk, and in that region estimates may be highly unreliable. Furthermore, a 'plateau' should not be overinterpreted as the probability of being 'cured' (see 'The proportional hazards (PH) assumption'). Notice in particular that if the patient with the longest follow-up has a failure (is not a censored case) then the Kaplan–Meier survival curve drops vertically to zero.

Hazard function

The hazard function $h(t)$ provides the 'instantaneous' risk of death at each time t for survivors at that time; that is, it represents the probability of dying at time t given that the patient survived up to that time and is also called the 'force of mortality'. We could say that while the survival function provides the clinician with the prognosis for a patient at the start of the disease history, the hazard function evaluates the risk of death continuously over time.

There is of course a relationship between the survival probability and the hazard function. Intuitively, the probability of surviving beyond t depends on how much risk you experience from the start to t. The following formula expresses this concept and introduces the 'cumulative hazard' $H(t)$:

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)du\right\}$$

As time progresses, the cumulated hazard increases and the survival probability decreases, but when the hazard is higher, the survival function decreases faster. We point out the fact that the probability of dying within time t depends uniquely on the hazard of death from 0 to t; the relevance of this remark will become clear when introducing the relevant objects in competing risks settings in the next section.

Two groups of patients are usually compared by means of the hazard ratio (HR):

$$HR(t) = \frac{h_A(t)}{h_B(t)}$$

HR = 1 indicates that the risk is the same; HR > 1 indicates that the risk is higher in group A, and conversely, HR < 1 indicates that the risk is lower in group A. The HR also provides a quantification of the difference, in percent. For example, HR = 1.2 means that in group A, the risk is 20% higher than in group B; HR = 2 means that in group A, the risk is double (100% higher) than in group B; and HR = 0.7 means that in group A, the risk is 30% lower than in group B. So, the HR provides a quantification of the effect of an exposure. However, in clinical terms it may be more relevant to look at the effect in terms of the difference of the survival probabilities $S_A(t) - S_B(t)$ instead of in terms of the HR.

The HR is the object of the analysis performed by the most common methods of comparing survival endpoints in two or more groups. In particular, this means the log-rank test (for marginal or unadjusted comparison) and the Cox regression (adjusted comparison, Section 4.3); both ideally require a PH assumption.

The standard methods for the analysis of survival are summarised in Table 2.

The PH assumption

The PH hypothesis has a key role in validating several methods of survival analysis, in particular the log-rank test and the Cox regression (Section 4.3).

Having the hazards proportional between two groups means that the HR is constant in time, and that the difference measured by the ratio between the two groups is always the same, at any point in time, from the beginning to the end of the disease history. It is also useful to remark that when applying the logarithmic transformation to the hazard function under the PH assumption, the difference measured in algebraic terms between two groups is equal to a constant, which means that the two log-hazard functions are parallel curves:

$$\begin{aligned} HR(t) = \frac{h_1(t)}{h_0(t)} \overset{PH}{=} \theta \Leftrightarrow h_1(t) = \theta \cdot h_0(t) \Leftrightarrow \log h_1(t) \\ = \log\theta + \log h_0(t) \end{aligned}$$

The same holds true for log-cumulative hazard functions. This property leads to a graphical method for assessing violations of PH, illustrated below.

This assumption has a specific implication in biological/clinical terms, and may not be satisfied in many situations. For example, characteristics or treatments associated with higher transplant-related mortality will act principally in the first months after transplant, and the HR will decreases to 1 after some time. Similarly, factors preventing relapse may show stronger effects in the long term, and thus HR values that diverge from the initial value and depart from 1. The extreme case is the combination of these situations, which leads to crossing hazard curves and a major violation of PH (a typical example in SCT is the comparison of allogeneic vs autologous transplantation).

The survival functions corresponding to two PH functions are neither parallel nor proportional, being related by $S_1(t) = S_0(t)^{\theta}$. If we have PH, then under the null hypothesis there is no difference between the two groups, $\theta = 1$ and the two survival functions are coincident at each point in time. This is why under an assumption of PH we can apply a test for the difference between two hazard functions (for example, the log-rank test) when we actually

[26] Standard deviations are usually computed applying the Greenwood formula. Note that the confidence intervals are obtained by neglecting the dependence between survival estimates at different times. For this reason, the band that connects the limits of the confidence intervals is not a proper confidence band for the whole curve. The R library km.ci implements several methods for computing pointwise confidence intervals as well as a 'simultaneous' confidence band.

want to compare the entire survival function. Nonetheless, it is important to stress that the log-rank test is a global test for the difference, thus even when the PH holds, a significant *P*-value from this test does not apply to the difference of survival curves at a specific point in time. A fortiori, in the presence of strong violations of the PH assumption, and in particular with crossing survival curves, the log-rank test is not useful to compare the curves, and actually any investigation on the global difference would be uninteresting. We will return to this issue shortly.

The PH hypothesis is important for the validity of the log-rank test because the test is most powerful in detecting a difference between two groups when this difference satisfies the PH assumption. It is less powerful or less capable of detecting an existent difference with a slight violation of proportionality such as converging curves (for example, HR decreasing to 1 after some time), and it generally fails in the case of strong violations, with crossing hazards, and crossing survival curves, because differences in the short and long term have opposite signs and tend to balance. The importance of PH in the Cox regression and those methods addressing non-PH in the framework of the Cox model will be illustrated in Section 4.3. In that section (especially in Section 4.3.5), we will also briefly discuss the possible causes of non-PH.

With converging survival curves, if the interest is mostly in the difference at the beginning of the follow-up, then there are tests belonging to the same family of the log-rank test that focus on early differences by assigning them more weight (Wilcoxon test and others).

When the primary interest is on the long-term outcome, even with PH and especially with non-PH and crossing survival curves, methods other than the log-rank test should be used. It has already been stressed that the log-rank test assesses a global difference and does not allow one to draw conclusions regarding specific time points or the long term in particular. It is also worth remarking that observing a 'plateau' in a Kaplan–Meier curve cannot be considered a proper method of assessing the probability of long-term failure-free survival or 'cure'.

When one is interested in the difference at a specific point in time, such as 5-year survival rates, then there are specific methods for computing the confidence intervals or testing the difference, including those proposed by Klein *et al.*[13] and Logan *et al.*[14] If the interest focuses on the probability of cure from the disease (that is, of having a failure rate similar to that of the general population, a type of endpoint very relevant in paediatric studies), one should move to cure models.[15,16] These topics are reviewed in the framework of adjusted comparisons (Section 4.3.5).

How to detect violations of proportionality

There are a number of methods described in survival analysis textbooks to detect violations of proportionality including in books by Klein and Moeschberger[17] (Chapter 11), Therneau and Grambsch[18] (Chapter 6), and Hosmer and Lemeshow[19] (Chapter 6). We mention here the two main approaches:

- Graphical check: Because of the relationship between $S(t)$ and $H(t)$, the log-minus-log transformation of $S(t)$ corresponds to log $H(t)$. Under the PH assumption, these transformations of the Kaplan–Meier curves for two or more groups should appear as parallel curves. This type of plot is easily produced by many software programs.
- The Cox regression: see Sections 4.3.3 and 4.3.4.

2.2.2. Competing risks endpoints

When the endpoint of interest is the occurrence of an event which has competing risks, as when relapse competes with NRM (Example 1), the relevant objects for the analysis are the cumulative incidence function $C_1(t)$ and the cause-specific hazard (CSH) function $h_1(t)$. Notice that we are using a subscript 1 to refer to the first of k competing events. In the case of relapse and NRM (which we will use for illustration), $k = 2$ and the subscripts 1 and 2 refer to relapse (as the main event of interest) and NRM (as the competing event), respectively. Of course there is symmetry, so every method is also applicable for investigating NRM, as well as in any other context of competing risks analysis.

Cumulative incidence

The cumulative incidence function $C_1(t)$ states for each time t the probability of having relapse before time t, or the percentage of patients who are expected to experience relapse within time t. At the time origin, $C_1(t) = 0$, and in time $C_1(t)$ increases until a total rate of relapse, which because there is a competing risk and the event of interest may not occur at all can be lower than 1 even after an 'infinite' follow-up.[27] Notice that the concept of a median time to relapse is not always well defined, as the cumulative incidence curve may not even reach the level 0.5 (50%).

Correspondingly, $C_2(t)$ provides the probability of dying without prior relapse before time t, or the NRM rate at time t. As it will be demonstrated below, the sum of the two cumulative incidence functions returns the overall probability of failure, regardless of the cause; the complement to 1 (100%) of this quantity is the failure-free survival probability (RFS in our example).

There is a specific nonparametric estimator (which has no particular name) for the cumulative incidence. A seriously mistaken approach that was very common in the past is to estimate this value as one minus the Kaplan–Meier estimate obtained by applying censoring to the cases failing for the competing event. This OMS-based approach returns a biased estimate of the cumulative incidence.[20] It is always overestimated, and the amount of overestimation increases with the importance, in terms of rate, of the competing event. In other words, the higher the incidence of NRM, the higher the overestimation of the incidence of relapse. What the OMS-based method returns represents a 'fictional' probability of relapsing if NRM risk could be removed completely and independently of relapse, that is, without changing the risk of relapse, which is clearly very far from reality. Unfortunately, the proper nonparametric estimator of the cumulative incidence is not available

[27] In probabilistic terms, this translates into the remark that $C_1(t)$ is not a cumulative distribution function; see, for example, Kalbfleisch and Prentice,[43] Chapter 8, and Marubini and Valsecchi,[2] Chapter 10.

among the standard methods implemented by many statistical software programs, although specific macros can be found. In R, the estimator is included in the library `cmprsk`.

For the assessment of differences among cumulative incidence curves, a number of established methods for unadjusted and adjusted comparison have been established, such as the Gray test[21] and the Fine and Gray[22] regression model,[28] respectively (both are implemented in R in the `cmprsk` library). Both of these methods have been criticised (for example, regarding their interpretation, see footnote 28), and alternative methods are now appearing in the current literature. In particular, adjusted cumulative incidence curves for different covariate patterns can also be estimated under certain assumptions by applying multi-state models (R library `mstate`) or by using methods based on pseudo-values[23] (SAS macros available).

Cause-specific hazard (CSH)

The CSH function $h_1(t)$ provides the 'instantaneous' risk of relapse at each time t for survivors without relapse at that time, or more generally, the conditional probability of failing for cause 1 at that point in time given that the patient did not fail for any cause before that time. It is worthwhile noting that the latter condition means no failure for relapse and no failure for NRM before time t. There is a symmetric meaning for $h_2(t)$.

Patients are at risk of both types of failure, and the sum of the two CSHs returns the overall hazard of failure regardless of cause (the hazard function for RFS). Thus, the survival function corresponding to the composite survival-like event (RFS) is:

$$\begin{aligned} S(t) &= \exp\left\{ -\left(\int_0^t h_1(u)du + \int_0^t h_2(u)du \right) \right\} \\ &= 1 - C_1(t) - C_2(t) \end{aligned}$$

This relationship between RFS and the sum of the cumulative incidence functions of relapse and NRM was anticipated above.

Because of the definition, in practice the inference for the CSH of an event of interest can be performed with the same methods used for the hazard function of a survival-like endpoint (Section 2.2.1) applying censoring to the observations failed for competing events. It is worth repeating that while this is formally correct within the use of the log-rank test or the Cox model for the analysis of the CSH, in general competing events cannot be treated by censoring (Sections 2.1.1, 2.1.2 and 2.1.7) because censoring indicates that the event of interest will occur after the date of last follow-up, while the occurrence of a competing event implies that the event of interest will not occur at all.

The standard methods for the analysis of competing risks are summarised in Table 3.

The relationship between cumulative incidence and CSH

In intuitive terms, the probability of experiencing relapse is expected to be higher with high instantaneous risk of relapse, but this is not the only element involved. In order to experience relapse, the patient also needs to avoid failing for NRM. Thus, we also expect the rate of relapse to be lower when the instantaneous risk of NRM is high. The following formula formally establishes the relationship between the cumulative incidence for the first event $C_1(t)$, the CSH for the same event $h_1(t)$, and the overall failure-free survival probability $S(t)$, which, as we demonstrated above, depends on the CSH of both the competing events:

$$C_1(t) = \int_0^t h_1(u)S(u)du$$

The practical consequence is that the results of the analysis of the cumulative incidence curves (for example, applying the Gray test for comparing two groups A and B) and the results of the analysis of the CSH (applying the log-rank test for the same comparison A vs B) may appear inconsistent; instead, both are correct, but look at different aspects of the same phenomenon. For example, if group A has a higher instantaneous risk of relapse among the patients alive in remission (a higher $h_1(t)$), but also has a higher risk of death without relapse (a higher $h_2(t)$), then the latter effect of A vs B may prevail and thus group A could have a higher NRM ($C_2(t)$) than group B and a lower relapse rate (a lower $C_1(t)$). Thus, A compared to B could be found to be a (significant) risk factor for relapse when looking at the CSH (by log-rank test or Cox regression), while when applying the Gray test or the Fine and Gray model, it may turn out to be non-significant or even significantly protective against relapse.

Should I analyse the cumulative incidence or the CSH?

The choice of what to analyse should not depend (only) on the availability of the software. The two objects correspond to two different perspectives.[24,25]

Looking at the cumulative incidence corresponds to focusing on the probability of having relapse, or on the percentage of patients who is expected to experience a relapse within a certain period after transplantation. This is the perspective of prediction from the start of the disease history, and it is relevant to informing the patients and making strategic decisions prior to transplant.

Studying the effects of factors on the CSH of relapse corresponds to investigating what factors increase the instantaneous risk of relapse among the survivors, information that is more useful from the researcher's point of view to investigate the biological mechanisms or for clinical decisions to be made after transplantation. However, it is a somewhat restricted analysis in the sense that it neglects to see the effects of the same factor on mortality. Thus, the analysis must look at both the CSHs of the competing events[29] and/or at the combined failure-free survival. In this respect, a good approach to reporting competing risks is by

[28] The Gray test and the Fine and Gray regression model work best under the assumption of the proportionality of a mathematical quantity called the 'sub-distribution hazard', which unfortunately has no clinical meaning. This makes it difficult to determine when this hypothesis of proportionality may not occur; see for example Scheike and Zhang.[45]

[29] An analysis of both CSH is efficaciously combined in a multi-state framework.

plotting stacked cumulative incidence curves for all competing events (Figure 2).

In addition to this type of reasoning, choosing a proper approach should also consider more technical issues. For example, proportionality cannot hold[26,27] for both a Cox model for the CSH $h_1(t)$ and for a Fine and Gray model for the cumulative incidence $C_1(t)$. Despite this limitation, performing both the analysis of the CSH and the analysis for cumulative incidence could be helpful in understanding the phenomenon from a wider perspective.

2.2.3. Analysing the occurrence of an event when the follow-up is complete

In SCT, there are several events of interest such as engraftment or acute GVHD (in its traditional definition[30]) that are evaluated at a fixed day x or within x days after transplant. Often the object of interest regarding these events is the overall rate of occurrence by time x, and not the exact timing.[31] In this case, if the follow-up is complete up to the day of evaluation for all or almost all patients such that for each patient the outcome is known, then the analysis can be based on simple percentages. The relevant statistical methods are briefly introduced here and are summarised in Table 4. Of course, if the follow-up is not complete at the time of the assessment x, or if there are event-free cases with last follow-up before time x, or if the interest is in the time of occurrence of the event, then proper methods for survival and competing risks analysis must be used.[32]

In this section we offer remarks on the definition of the outcome as dichotomous or categorical variables with $k \geqslant 3$ levels. In fact, regardless of the particular event you are interested in, death is usually a competing risk (unless the event is death itself). In principle, then, the outcome variable is a categorical one with at least three levels that we can code as 0 for patients who are alive and event-free at day x, 1 for patients who experienced the event of interest within time x, and 2 for those who failed for a competing event within time x. More levels (coded 3, 4 and so on) could be necessary to distinguish among different types of competing risks, if this is relevant for the investigator. Notice that all cases with events occurring after time x must be coded as zeros for the analysis.

In practice or in some clinical settings, early death could be an irrelevant competing risk in terms of incidence, or else the investigator may not be interested in distinguishing between failure-free cases and cases failed for competing events. In these situations, the outcome variable is a dichotomous one, indicating the occurrence or not (1 / 0) of the event of interest, and it is summarised by a simple percentage (the number of cases who had the event before time x divided by the total number of cases evaluated) which estimates the probability π that a patient experiences the event before time x. Usually two groups, A and B, are compared by the risk ratio π_A/π_B. Significance can be tested in various ways, for example by χ^2 or Fisher's exact test applied on the 2-by-2 contingency table (group by event). The adjusted comparison can be calculated using the logistic regression, where the effects are estimated in terms of the odds ratio:

$$OR_{\text{AvsB}} = \frac{\Omega_A}{\Omega_B} \quad \text{being } \Omega = \frac{\pi}{1-\pi}$$

When it is more appropriate to distinguish event-free cases from cases that failed for competing causes before experiencing the event of interest, or to separately report different types of competing failure, the tools for a descriptive analysis of the outcome are the same: a table of percentages for each of the k levels and a χ^2-test on the 2-by-k table. Some care is required in the communication of results, as is suggested in the example below. However, the multivariate analysis is more complicated (methods such as an extension of logistic regression[33] can be applied).

Example 4: Engraftment rates

Haematopoietic recovery (or engraftment) after transplantation corresponds in general terms to the achievement of persistent blood cell counts above predefined levels, usually evaluated within 30, 60 and 100 days from transplant. (The precise definition depends on specific clinical/biological issues that may differ with respect to the type of transplant or the source of stem cells. The official EBMT definitions should be used in EBMT studies).

Early death is a competing risk. Other competing failures could be graft failure, loss of graft, second transplant with no prior engraftment, and even relapse or disease progression. The statistician should discuss with the clinical investigators the role of these events, although there is no great necessity for taking them into account if they have very small incidence. The analysis should consider all relevant competing events.

Consider a situation where engraftment is evaluated at day 30 after transplant. The main interest is the overall rate (probability) of engraftment, and the only competing risk is death without engraftment. The outcome is thus a three-level categorical variable. Out of 10 patients, two died without prior engraftment at days 10 and 14, respectively, and the others engrafted at days 5, 9, 11, 12, 15, 25, 35 and 45. All cases can be evaluated at day 30, that is, there is no censoring. Because this is true, we can look at percentages instead of cumulative incidence curves. Notice that of the

[30] The 'traditional' separation of GVHD into acute and chronic based on a time threshold (100 days) is currently being replaced on a clinical basis, and as soon as data collection rises above this rigid definition, the statistical analysis will necessarily switch entirely to the competing risks setting. To date, acute GVHD was most often analysed as described in this section, while chronic GVHD as a competing risks endpoint was left-truncated (Section 2.1.4) at 100 days, thus restricting to survivors at 100 days.

[31] The timing of engraftment, for example, would be described as the median among the cases who had engraftment.

[32] For an event without competing risks the methods proposed in Klein *et al.*[13] and Logan *et al.*[14] may be used to compare the overall rates of events at time x even with censored cases.

[33] Indicate with E1 the event of interest and with E2 the competing event. The corresponding probabilities of occurrence are π_1 and π_2; $1-\pi_1-\pi_2$ represents the probability of being alive and failure-free (no event) at the time of assessment. Correspondingly, define:

$$\Omega^1 = \frac{\pi_1}{1-\pi_1-\pi_2} = \frac{\Pr(E_1)}{\Pr(\text{noevent})} \quad \text{and} \quad \Omega^2 = \frac{\pi_2}{1-\pi_1-\pi_2} = \frac{\Pr(E_2)}{\Pr(\text{noevent})}$$

To compare two exposure groups, A and B, it is possible to implement regression models, where the effect of A vs B is estimated in terms of the ratios:

$$OR^1 = \frac{\Omega^1_A}{\Omega^1_B} \quad \text{and} \quad OR^2 = \frac{\Omega^2_A}{\Omega^2_B}$$

engraftments observed, only six occur before day 30, meaning there are two event-free patients. It is not complete reporting only that '6 out of 10 (60%) engrafted', it is more appropriate to provide the full information: '6 out of 10 (60%) engrafted before day 30, and another 2 patients (20%) died without prior engraftment. The remaining 20% of the patients were alive without engraftment at day 30.' With this information, the reader realises that another 20% of cases may have engrafted after day 30, thus the current 60% probability may have risen to 80% at a later time (assuming, of course, that this is clinically feasible).

2.3. Population selection and methodological issues

The study population must be defined according to the rationale of the investigation and must represent a target general population of interest. A limited possibility for generalising the results of the analysis and the presence of bias, that is, some inbuilt systematic 'error' that leads to faulty knowledge, are the main problems to be aware of when selecting the study population. In this section, we present examples of biased selections (not all possibilities are covered).

In RBS, the study population is selected according to the rationale of the investigation by including only certain types or subtypes of disease, restricting to certain patients or transplant characteristics, fixing the calendar period and so forth. It is worth pointing out that when the endpoint of interest is survival since first transplant, for example, the selection criteria cannot be based on patient characteristics defined after first transplant, nor on future events (such as the administration of a second transplant) or outcomes. The original selection can be refined during the phase of preliminary descriptive analysis. A main reason for doing so could arise from the presence of missing values. The problems that may arise and how to conduct this delicate phase are illustrated in Section 3.3. The following examples illustrate bias arising from selecting the population to exclude missing values. Missing values do not always cause problems, and neither does bias originate only from missing data.

- An important risk factor (for example, a biomarker) is in principle recorded on EBMT Med-B forms, although it is more frequently reported by large, experienced transplant centres. Selecting only cases with non-missing information for this factor returns a study population that may not be representative of the general target population; such reference centres may treat the 'worse' patients or they may provide better care and thus better outcomes than the 'average' centre. Notice that the investigator may be unaware of this 'hidden' selection mechanism because it is difficult or nearly impossible to identify centre characteristics and/or relate them to missing values.
- Often missing values are related to the calendar: some risk factors have only recently been identified and were not or were only rarely collected in previous years, some treatments were in use during certain periods, but seldom reported in other periods, and so forth. Thus, restricting the analysis to cases with known information can correspond to a selection based on other phenomena associated with calendar time: improvements in treatments, changes in the diagnosis of relapse, admission to transplantation of wider categories of patients and so on.
- If missing values more frequently affect one of the main subgroups whose comparison is the target of the study, then a selection of known cases could induce a bias, as in the following example. A study aims at comparing two diseases, A and B. Cytogenetics is considered an important adjustment factor and thus only patients with known cytogenetics are included in the analysis. Unfortunately, in the past cytogenetics were usually performed for patients with disease A, while in patients with disease B it was conducted only if some other risk factor was present. Thus, after the selection group B includes the 'worse' patients, and this bias affects the comparison with group A. (The analysis should at least try to control for the effect of the third risk factor, but this may not be sufficient to correct for this bias).

In prospective studies (both observational and interventional), the definition of the study population is made during the planning phase by fixing inclusion and exclusion criteria for enrolment (and possibly defining sub-populations for specific analyses) and in general should not be refined after seeing the data. In particular, the efficacy comparisons should be performed by analysing all cases enrolled and applying the ITT principle.

This important concept applies to any study comparing treatments, say A vs B, where the decisions on which treatment should be given to each patient was determined according to a protocol or otherwise (for example, in RBS) recorded at the beginning of the disease history. The clearest situation is the case of a randomised clinical trial. Following the ITT principle, we basically compare the outcomes between the groups defined on the basis of the treatment assigned, regardless of subsequent events (and of actual treatment). In particular, each observation is included in the analysis and belongs to the original group, A or B, regardless of whether the patient had a treatment schedule or dosage modified, failed to complete the treatment or failed to receive the prescribed treatment at all (by, for example, crossing-over to the other arm). The rationale is that the need for modifying, stopping or changing the treatment could be related to treatment, and failing compliance could be interpreted as a failure of the treatment strategy. Reasons for non-compliance are often related to toxicity or more generally to secondary effects of the treatment, including consequences of particularly uncomfortable or heavy treatments such as depression or discouragement (and thus drop-off for refusal or loss to follow-up) and even suicide. Seen from another perspective, the efficacy of a treatment may allow the patient to receive further treatments not originally prescribed (for example, a second transplant) that provide a benefit for the final outcome. Excluding from the analysis or re-allocating patients according to the treatment actually received (the criterion known as per-protocol, PP) is a potential source of bias except when non-compliance was completely independent of the status of the patient. The following example illustrates a bias arising from violating the ITT principle in the analysis of data from a PCT.

- A PCT aims to compare two treatments, A and B; the treatment arm is assigned at enrolment according to randomisation. At the end of the study, many patients turn out to be non-compliant, going off treatment well before they received all the planned cycles.[34] Following the feeling that for these cases the treatment could not perform its curative action, investigators exclude them from the analysis. Because non-compliance was mostly due to toxicity, and treatment A was more toxic than treatment B, the PP population is self-selected; in particular, group A remained the fittest patients who had fewer problems with toxicity or could otherwise overcome them. The PP comparison between A and B is thus biased and overestimates the efficacy of treatment A. This PP estimate is a measure of the effect of A provided that the patient will not interrupt the treatment. It is not really useful for deciding how to treat a patient because, in real life, patients may be non-compliant. It is worth noticing also that while the ITT groups A and B were created by randomisation and, thus, were comparable in terms of baseline characteristics, the PP groups have lost the randomisation advantage.

The bias caused by self-selection mechanisms like the one described above may also affect the RBS, where they may be less clearly identifiable and where perhaps it is less feasible to find a correct approach. We have already introduced an example of potential self-selection in a situation with second transplantation (Section 2.1.6). In the EBMT registry, information on the planned treatment strategy at each transplant is, in principle, available, which can be used to make an ITT analysis to compare different strategies, such as autologous plus allogeneic vs double autologous transplantation. However, because of the retrospective data collection, this type of analysis is not comparable to a randomised study, thus necessitating very cautious interpretation.

3. Preliminary, descriptive and marginal analyses

The general scope and features of preliminary, descriptive and marginal analyses were introduced in Section 1.3 together with the problem of controlling for confounding factors and, more generally, the need for the simultaneous evaluation of the effects of several prognostic factors on outcomes. The latter phase of the statistical analysis is treated separately in Chapter 4 (adjusted comparisons). The main methods used for preliminary, descriptive and marginal analyses are indicated in Section 3.1.

The preliminary phase is critical to the development of the study because it is the phase that fixes the data (Section 3.2) and refines the study population (in RBS) and sometimes the entire plan of analysis. In this phase, it is crucial to investigate potential problems related to the presence of missing values (Section 3.3). Another important aspect, evaluating the actual amount of information available from the data on the occurrence of the events of interest, is treated in Section 3.4.

Marginal or 'univariate' analysis (the analysis of each endpoint of interest with respect to one or more factors, separately for each factor and with significance testing) is valuable in itself as part of the descriptive analysis, but usually the conclusions reached must be confirmed by a multivariate analysis (adjusted comparisons are treated in Chapter 4). The marginal analysis is thus an intermediate and fundamental step of the adjusted analysis. At this stage, multiple tests are being performed, and the probability of the inflation of false discovery is non-negligible. To prevent false conclusions it is possible to adopt rules such as the Bonferroni–Holm correction (illustrated in Section 4.2.3), and apart from applying these rules, it is important to appraise significance in an appropriate way (Section 1.4) when interpreting the results.

Given the importance of an objective communication of results, this chapter also presents information on how to report tables and curves (Section 3.5).

3.1. Methods for descriptive and marginal analyses

Preliminary, descriptive and marginal (that is, unadjusted) analyses can be performed using the methods indicated in this section. Table 1 lists the basic statistical methods that can be used to analyse variables (patient characteristics, transplant type, and so on) to check and describe them, and perform marginal hypothesis testing. We recommend consulting standard statistics textbooks for broader illustrations of these and other suitable methods, as well as for what factors influence the validity of these methods and how to test for it (for example, how to test normality and equality of variances before applying a *t*-test). The methods for describing and marginally comparing endpoint variables were illustrated in Section 2.2, and are now summarised in Tables 2–4. The estimation of median follow-up is described in Section 3.4.

Table 1. Methods for descriptive statistics

		Test for association	
Type of variable	Description	Differences in k groups	With a continuous variable
Quantitative continuous Example: WBC; age	Median and other quantiles; minimum and maximum value	Mann–Whitney test ($k=2$) or Kruskal–Wallis test ($k>2$)	Test on the correlation coefficient (linear association); Spearman's Rho, Kendall's Tau-tests (more general association)
- if with normal distribution (symmetric, bell-shaped), also:	Mean and standard deviation	*t*-Test ($k=2$) or analysis of variance (ANOVA) ($k>2$)	Linear regression
		These methods require verifying certain hypotheses, for example, homoschedasticity	

[34] It was already remarked that censoring observations when the patients go off-protocol is erroneous (Section 2.1.1), and it is suggested in this case to define combined endpoints to account for interruptions or violations of the prescribed treatment, considering them as types of failure (Section 2.1.3).

Table 1. Continued

Type of variable	Description	Test for association	
		Differences in k groups	With a continuous variable
Categorical k levels Example: CML Phase $= \{CP, AP, BC\}$ ($k=3$ levels); gender $= \{M, F\}$ ($k=2$, dichotomous). Example of ordered categorical: stage of the disease (I–IV); age in classes	Frequency table	χ^2-Test or Fisher exact test (2×2 tables; preferable for small samples) on the cross-tabulation (specific tests of trend can be applied for ordered variables)	(Same: Mann–Whitney, Kruskal–Wallis, t-test, ANOVA)

Table 2. Methods for survival endpoints (for example, OS, RFS, PFS)

Object	Estimation	Unadjusted comparison	Regression model
Survival function $S(t)$	Kaplan–Meier curve		
Hazard function $h(t)$		Log-rank test* Wilcoxon test	Cox* (estimates of the effect via hazard ratios)

*Methods with the best performance when PH hold.

Table 3. Methods for competing risks endpoints (for example, relapse, NRM, chronic GVHD, response)

Object	Estimation	Unadjusted comparison	Regression model
Cumulative incidence $C_j(t)$	Proper non-parametric estimator*	Gray test**	Fine and Gray** (effect estimates: no intuitive clinical meaning!)
Cause-specific hazard function $h_j(t)$		Log-rank test***	Cox*** (effect estimates: cause-specific HR)

* Do not use OMS from a Kaplan–Meier curve (see Section 2.2.2).
** Methods with the best performance when the proportionality of the sub-distribution hazard function holds.
*** Methods with the best performance when the proportionality of the cause-specific hazard function holds.

Table 4: Methods for events before a certain time (complete follow-up) (for example, engraftment at day 30, acute GVHD)

Object	Estimation	Unadjusted comparison	Regression model
Probability	Percentages	Tests for frequency tables (χ^2 or Fisher exact for 2×2 tables) or specific tests for probabilities/ risk ratio/ odds ratio	Two levels: logistic regression (effect estimates: odds ratios) (three levels: for example, extensions of logistic regression)

The completeness of follow-up is fundamental. No patient should be lost to follow-up event-free before the end of the evaluation period. If this condition does not hold, use the methods in Tables 2 and 3.

3.2. Data preparation

The preparation of the final data set includes verifying nonsense or inconsistent values, extreme values, missing data and performing some preliminary recoding of the relevant variables, although decisions on this type of data management must be made in close association with other decisions encompassed by the final statistical analysis (for example, with regard to the treatment of missing values and the transformation of the variables).

Errors in the data. Despite data verification at the database level, data can still be affected by several errors, and preliminary verification will avoid losing time later following the discovery of problems in the data set during the analysis. Two main data errors may arise from nonsense values (for example, negative values for the time interval between diagnosis and transplantation) or inconsistent values (for example, a time to relapse that is greater than the survival time).

Tables and bar charts with frequencies (for categorical variables) or descriptive indexes (minimum and maximum values, 5% and 95% percentiles, the five lowest and largest values), and graphs (histograms and boxplots) for continuous variables can be used for a general overview. In addition, targeted assessments could be made based on the substantive clinical knowledge (for example, standard reference values for a clinical parameter at diagnosis) or on the basis of previous experiences with data from that particular registry or disease. All nonsense and inconsistent values should be corrected whenever possible; otherwise, they should be set equal to missing.

Extreme values. Values that are plausible (that is, those that cannot be considered nonsense) but that are quite distant from the other values observed in the sample (for example, more than 3 times the standard deviation in normally distributed variables) are called outliers. When these values can be attributed to data collection errors (during data entry, typing or conversion; sometimes they are caused by using the wrong measurement unit), they should be corrected or treated as missing values. Otherwise, they deserve special attention. On the one hand, they could indicate some biological mechanism that requires further investigation. On the other hand, they could drive the conclusions of the analysis, especially when dealing with small samples. For these reasons, during data checking outliers should be identified for subsequent investigations (such as influence analysis in multivariate regression models, mentioned in Section 4.2.6).

Missing values[35] usually affect both registry data and data collected from clinical trials.[36] Depending on the amount of missing data and on the precise nature of the missing data, important limitations or biases may affect the analysis and its conclusions. It is therefore crucial during the preliminary phase to search for missing values and reduce them if possible by retrieving the data or imputing the correct value if it can be derived from other information (such as when for historical reasons an item was not applicable or deterministic, for example, the type of conditioning could not be 'reduced' in 1985). Investigating the potential impact of missing data on the analysis and deciding how to manage missing data are also part of the preliminary analysis and can result in the refinement of the study population. This is a delicate phase of the study and may require the contributions of expert statisticians and principal investigators. Section 3.3 is dedicated to missing values.

Data transformations. Preliminary data transformations (apart from those foreseen in the study plan) can appear necessary after the preliminary descriptions have been made. Continuous variables with skewed distributions may require a log-transformation (to reduce the influence of very high values) or another functional transformation. Qualitative variables (or discrete numeric variables that can assume only few values) could be recoded by collapsing categories when the number of patients in some of the original categories is low. Of course, the categories being created will have to retain proper clinical meaning (consistent with the aims of the study), and it is strongly recommended to retain properly ordered categories for variables such as disease stage. Further data transformations may be applied in building regression models (see Sections 4.2.2 and 4.2.3).

3.3. Missing values°

Any statistical procedure (from a simple frequency table to a regression model) that involves one or more variables with missing values can be performed only by excluding all cases with one or more of the missing values. The first consequence of having missing values, therefore, is that the analysis is based on a restricted population: estimates are less precise, and there is a loss of power for statistical hypothesis testing. Notice also that different procedures involving different sets of variables will each refer to a different (sub)population. Problems are limited to the loss of sample size only in the case of 'missing completely at random', where there is no pattern of relationship between the missing values and any known or unknown influential characteristic or outcome (including the variable itself). In this situation, the sub-populations are representative of the target population.

When the missing values are differently distributed according to other factors, or when they may possibly be related to unknown or unmeasured characteristics, and the cases with known values have different outcomes than the cases with missing values, then the conclusions drawn from an analysis on a restricted sub-population may be affected by bias (and, if each procedure excludes a different set of observations, every analysis may be biased in a different way). The problem is partially amendable if it can be assumed that the nature of the missing data is related to known characteristics (missing at random, MAR), both because some 'good' technique of imputation (Section 3.3.2) allows predicting the missing value from other information present and because the results can be interpreted. If missing values are truly related to an unobserved variable that affects the outcome, or depend on the actual value (for example, large values of X are more frequently missing than small values), then imputation techniques are not useful, and the conclusions of the analysis are questionable (this is an informative missing, IM, problem).

It is therefore important to investigate (Section 3.3.1) the amount and type of missing values and to understand their potential impact on the analysis and its conclusions to decide how to treat missing values. The results of the analysis of the type of the missing data should also be described when publishing the results to allow readers to evaluate the potential limitations of the study and to support the appropriateness of the statistical analysis performed.

As a general rule, the reduction of the study population to the cases with known values for a set of key variables can be considered provided that the amount of missing data per variable and globally does not exceed 5–10%, and/or that the resulting sample size is still adequate to the study objectives and there is no evidence of relevant bias. Alternatively, it may be better to disregard a variable with many missing values when performing the multivariate analysis.

A further option, at the cost of more complicated and time-consuming analysis, is the application of methods for the statistical imputation of missing values. Be aware, however,

[35] We refer here to explanatory variables. Cases with a missing value for the main endpoints are usually excluded from the study, although the investigators should always be concerned with generalisation and bias (Section 2.3) and undertake the necessary investigations. A few missing values in secondary outcomes may be tolerated, but the reduction of the sample for that particular analysis must be documented in the publication, together with some discussion on whether the restricted population is still representative of the whole sample and whether there is any possible relationship between missing data and the value of the endpoint. The reasons are explained in the text.

[36] In clinical trials, the presence of missing values for relevant variables is particularly problematic (for issues related to the ITT principle, see Section 2.3). The strategy used to manage missing values should always be indicated in the protocol of any type of prospective study. Avoiding missing values as much as possible should be planned for during any data collection or prospective study. Usually, it is necessary to restrict data collection to the main items needed, implement a good method of data entry, perform monitoring to avoid errors and so on.

that no method can really correct for hidden or unclear mechanisms related to the nature of the missing data. In any study, the 'best' solution, although expensive in terms of time and budget, is to improve the quality of the data.

3.3.1. Investigating the nature of the missing data in practice

For a single variable X, or when assessing the consequences of the exclusion of cases with missing values for a set of variables, a first step is to determine whether the nature of the missing data is related to any of the main outcomes or to other characteristics. The deletion of missing values can be considered 'safe' with respect to potential bias if the 'missing' group has outcomes similar to the 'known' group and presents similar characteristics. This is confirmed by defining and comparing the 'missing' and 'known' subgroups; the comparison can be made using univariate methods (Tables 1–4) or multivariate analyses (for assessing relationships to certain variables, the choice may not be restricted to logistic regression, but may include cluster analysis, recursive partitioning, and so on).

A slightly different approach is to analyse the missing X as an additional category of X:[37] the outcome of the 'missing' level should be intermediate among the outcomes of the other levels. For example, high Beta2 values are a risk factor, and myeloma patients with missing Beta2 are expected to have an estimated risk of death higher than that of the group with 'low Beta2' and lower than the risk of the group with 'high Beta2'. Moreover, if the 'missing' subpopulation is similar to the general population, and if the latter has a majority of cases with low Beta2, then the risk estimate for the missing group will be closer to the estimate for low Beta2. In some situations, the use of the 'missing' category for a variable X to be included in a model may be an acceptable alternative to the deletion of cases with missing values.

Unsatisfactory, unexpected and inexplicable results from these simple analyses deserve additional investigation and indicate caution in the management of missing values.

3.3.2. Statistical imputation methods*

Imputation is an alternative to case deletion. In general terms, it consists of replacing the missing values with values chosen from the range of possible values on the basis of the available data. The advantage with respect to case deletion is efficiency in that no information is wasted, but rather all information is used to 'predict' the missing values. The disadvantages come from relying on assumptions and the practical difficulty.

'Single imputation' methods use a single 'best guess' for each missing value, and perform the statistical analysis on the resulting complete (imputed) data set. 'Multiple imputation' methods impute more than one value to each missing value, then perform the analysis on each complete data set, finally returning as a result some average of the results from each imputed sample (adjusting the variance and covariance matrix).

A very simple and traditional single imputation approach is to replace a missing value with the median of the observed values (X continuous covariate) or with the mode (X categorical covariate). This type of approach could be considered when there is a small percentage of missing cases for very few variables. A more effective approach is to use regression models estimated on subsets of the known cases to predict the missing value of X_1 from other variables: X_2, X_3, and so on. The literature proposes several different methods for this type of imputation (for example, many authors suggest imputing the value estimated from the model corrected by a random component, including the possibility of sampling from the residuals of the models, and so on). A further approach is matching a 'missing' with a 'known' value, but this may be unfeasible unless the sample is very large. One important point is that the most recent statistical literature recommends including the outcome among the predictors of missing values; otherwise the relationship between the variable with missing values and the outcome would be underestimated.

This quick overview is only meant to recommend avoiding a 'home-made' approach to handling missing data when there is extensive literature illustrating many methods for dealing with the problem. A good review can be found in Chapter 3 of the book by Harrell,[28] which also illustrates the use of the R library `Hmisc` for some of the methods. It is clear that more 'sophisticated' techniques are more difficult to implement; however, they are more appropriate for complicated situations where case deletion is not really an option (in the presence of high percentages of missing values in many relevant covariates, correlations among missing data and known information, and so on). The major statistical programs implement one or more imputation techniques; it is thus recommended that the statistician assesses what is being used, makes a cautious use of these methods, and remains aware that no method can truly correct for hidden or unclear mechanisms related to missing data. Clues of such mechanisms are provided by exploratory analysis, the comparison of results of different imputation approaches and by unexpected results in general.

3.4. The length of follow-up

Usually, studies of long-term (time-to-event) outcomes must report the length of follow-up in terms of the median (with 95% confidence interval). There are two methods for computing the median follow-up. One is to use the median estimated from a ('reversed') Kaplan–Meier curve for the survival times of all patients, where deaths are censored (code = 0) and failure-free patients provide complete observation (code = 1). The rationale is that death prevented observation of the actual follow-up time. Another method is to compute the median of the survival times only from patients alive at last follow-up; however, this estimator may be unsatisfactory with high failure rates.[29] Another object of interest could be the difference between the potential follow-up (time from origin to cutoff date of analysis) and the actual follow-up for failure-free cases.

[37] This holds for X categorical covariates; if X is continuous, the best approach is to categorise it (in some texts, it is suggested instead to replace the missing value with a fixed value such as the mean or median of the known values of X, and add a dummy variable to indicate missing data (1 if missing, 0 if known) in the models (dummy variables are illustrated in Section 4.2.3).

3.4.1. Follow-up, events and censoring: do I have enough information?°

In a study of survival, a major concern is whether the follow-up and the information that was captured are 'sufficient' to draw conclusions. There are no precise answers, but some remarks are appropriate.

Technically speaking, the amount of available information for a survival analysis is measured by the number of failure events observed, not by the total number of cases analysed. Thus, one could say that the follow-up is 'short' when the number of events observed is small. In particular, the statistician may observe that the number of events observed does not allow a statistical evaluation of the objects of interest; for example, as a rule of thumb, one should have observed 7–10 events per parameter to be estimated in a Cox regression.

From a different perspective, a study of survival times must have a follow-up adequate with respect to the expected timing of failures. The clinical investigators may judge that a large percentage of censoring indicates that the follow-up was insufficient with respect to the 'speed' of failure for that disease/group/treatment.

If the potential follow-up of the study appears to be adequate for its purposes, observing few events may indicate the unfortunate possibility that failures are under-reported. This leads to concerns about IM and biased estimators (Section 2.1.1) and the validity of the study. When under-reporting is suspected especially in the long-term, investigators could decide to artificially reduce the length of follow-up for the study. This approach can also be followed when comparing two groups with very different follow-up times (and high censoring rates), which limits the possibility that the role of that prognostic feature is influenced by the unequal observation pattern. Artificial censoring at time x is applied simply by recoding the survival indicator to be equal to zero when the failure time is larger than x.

Of course, when it is expected to see 'cured' patients, or when patients can have a large probability of being failure-free (or more precisely, they are at no more risk of failure than the general population) for a long time, as in some paediatric studies, large percentages of censored data do not raise concerns about the validity of the study, but the analysis may require specific methods (see 'The PH assumption' and Section 4.3.5 for cure models[15,16]).

3.5. Reporting tables and curves

3.5.1. Tables.

Tables are very useful synthetic prospects of the data. They are commonly used to report population characteristics (possibly in subgroups), survival or cumulative incidence probabilities, results from regression models and so forth. Below are a few suggestions for table contents:

- When describing the population in subgroups, always report the percentages computed on the proper denominator. For example, in a study including autologous and allogeneic transplants, the percentages for donor gender should be computed on the total number of allogeneic transplants, not on the total number of transplants included in the analysis. As another example, percentages for causes of death should be computed on the number of patients dead.
- Particular attention should be given to the presence of missing values. Computing percentages on the number of known observations (that is, excluding the missing values) is usually the most appropriate approach unless an interpretation can be given to the missing values (in particular, records such as 'not applicable' or 'not done' should be discussed with the clinical investigator and/or study coordinator).
- The number or percentage of missing values should always be reported.
- The sum of the percentages should of course be 100% (round the percentages if necessary).
- In tables in particular, and possibly throughout the manuscript, round all quantities of the same type to an equal number of decimal figures.
    - ○ For *P*-values the journals often provide specific guidelines; as a general rule they should be reported with three digits after the decimal separator; for very small values it is suggested to choose '<0.001', while high *P*-values (for example, larger than 0.10) could be reported with only two digits after the decimal point.
- Whenever possible, when reporting estimates for subgroups to be compared or the effect of a factor from a regression model, report the confidence intervals in addition to or in place of the *P*-values (Section 1.4).
- The caption of the table should describe the contents of the table, the population on which it is based, and possibly the statistical methods used.
    - ○ Example: 'HRs (with *P*-values and 95% confidence intervals) from the adjusted Cox regression, including treatment, age and status at conditioning. All cases.'

Example of table reporting characteristics

Patient characteristics	Missing values (%)		n	%
>45 years age at diagnosis	0 (0.0)	No	48	13.4
		Yes	309	86.6
MM classification	6 (1.7)	IgG	210	59.8
		IgA	63	18.0
		Light chain	60	17.1
		Other Ig	6	1.7
		Non-secretory	12	3.4
β2 at diagnosis (mg/dL)	79 (22.0)	≤4	210	75.5
		>4	68	24.5

Notice that percentages in each subgroup are computed on the basis of the 'known' cases, not on the overall total (357). In this table, a column is dedicated to providing details on missing values, but these could alternatively be provided as a note below the table.

3.5.2. Curves

As was seen in Section 2.2, the Kaplan–Meier curves and the cumulative incidence curves (obtained from the proper nonparametric estimator) summarise survival-like and competing risks endpoints, respectively. A few suggestions for the presentation of these curves are given below:

- The range for the vertical axis (ordinate), representing the probability, should always be (0.0–1.0) or (0–100%). The scale for the horizontal axis (abscissa), representing time, should mark relevant time points from the time origin, whether in days, months or years.
- The time range should be restricted to avoid showing the tails of the curves where in one or more subgroups there are fewer than 5 or 10 patients still at risk. The estimates in these regions are highly uncertain, and showing long horizontal tails ('plateaus') based on few observations could improperly suggest an interpretation in terms of the probability of being 'cured'.
- More generally, the graph should allow the reader to appraise the precision of the estimates. The initial size of the population and the number of patients lost to follow-up must be clearly reported in each graph. Tick marks on the curves are used to indicate the censored cases, but this is not sufficient with respect to loss-to-follow-up. It is recommended that the number of patients at risk in each group for a series of times (below the time axis, as shown in Figure 5, left side) be reported.
- The precision of the estimates could in principle be reported by adding bands around the curves that represent the uncertainty of the estimates (Figure 5, right side). Notice that the majority of software programs can graph bands obtained by connecting the upper and lower limits of the pointwise confidence intervals (that is, the confidence interval (CI) for each point estimate). More appropriately (and difficult to do in practice), the variability of the entire curve should be evaluated by taking into account the dependence among all of the estimates. Such a confidence band is usually larger than the band based on the pointwise CIs. However, plots reporting bands are often not very clear, especially when the graph compares two or more groups.
- When reporting competing risks, consider using a graph of stacked cumulative incidence curves (Figure 2). The curves for all competing events can be plotted one on top of the other so that the area between the two curves (or between the first curve and the horizontal axis) reproduces the probability for each type of failure, and the area above the curve on top represents the total failure-free survival probability (Section 2.2.2).
- The caption of the plot should clearly indicate the endpoint (entry and exit times, failures/competing events being considered), the population (selection criteria) and in some cases (such as when reporting clinical trials) the date of the analysis.
    - ○ Example: 'PFS from transplant to either relapse/progression or death (or last follow-up), all cases.'

4. Methods for adjusted comparisons

4.1. Regression models

The general idea of a regression model is to describe how an outcome variable depends on a series of 'explanatory' variables considered together and not separately as in the 'marginal' analysis. For this type of analysis, the adjectives 'multi-variable' and 'multivariate' are used. The purpose is to estimate the 'net' effects of each explanatory variable and to control for confounding (Section 1.3).

More specifically, a linear regression model assumes that some quantity that identifies the probability distribution of the outcome depends on the explicative variables (say X_1, X_2 and X_3) through a linear combination $\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3$ (the term β_0, called the intercept, is not always included, see for example the Cox model). Thus, the regression parameter β_1 quantifies the effect on the outcome of a unity increase of X_1 given that X_2 and X_3 remain constant (this is the 'adjusted' effect of X_1). The beta parameters are estimated from the data. A test for the null hypothesis of no effect (H_0: $\beta_j=0$) is performed for each covariate; in addition, an overall test for the presence of any effect is provided.[38]

The best known regression technique is the multiple linear regression method, which applies to continuous, non-censored and normally distributed dependent variables. As was seen in Chapter 2, this type of outcome variable is hardly ever present in the context of SCT research where the main types of endpoints require the Cox model, the Fine and Gray model, and logistic regression (Section 2.2). Apart from some technical issues, the general remarks on model building are similar, and they will be illustrated in the next section, while Section 4.3 focuses on the Cox regression. As the title of this chapter implies, we will focus on the situation where the model is aimed at estimating the effect of one main factor, such as treatment, while adjusting for other variables. This objective is different from building a model for the prediction of outcomes or for proposing a risk score, situations where other issues (explained variation, calibration, and so on) should be taken into account and where other tools (such as regression trees) could be used. Providing detailed suggestions on building prognostic models properly is beyond the scope of these Guidelines; for a good review, see Harrel.[28]

It may be superfluous, but still fundamental, to remark that a 'true model' does not exist but that all models are only useful or not useful, and they are faithful or unfaithful to the data. Models are a type of 'simplification' of extremely complicated, highly heterogeneous things that apply some (mathematical) reduction of complexity to highlight some aspects of it. Several different models may be built for the same outcome and the same set of explanatory variables. Usually, we propose one single model, choosing the one that is more useful to focus on the object of interest in our study. What guarantees the

[38] The overall significance is tested using the Likelihood Ratio test. The Score test and the Wald test are based on approximations and generally lead to the same conclusions (if not, refer to the Likelihood Ratio test).

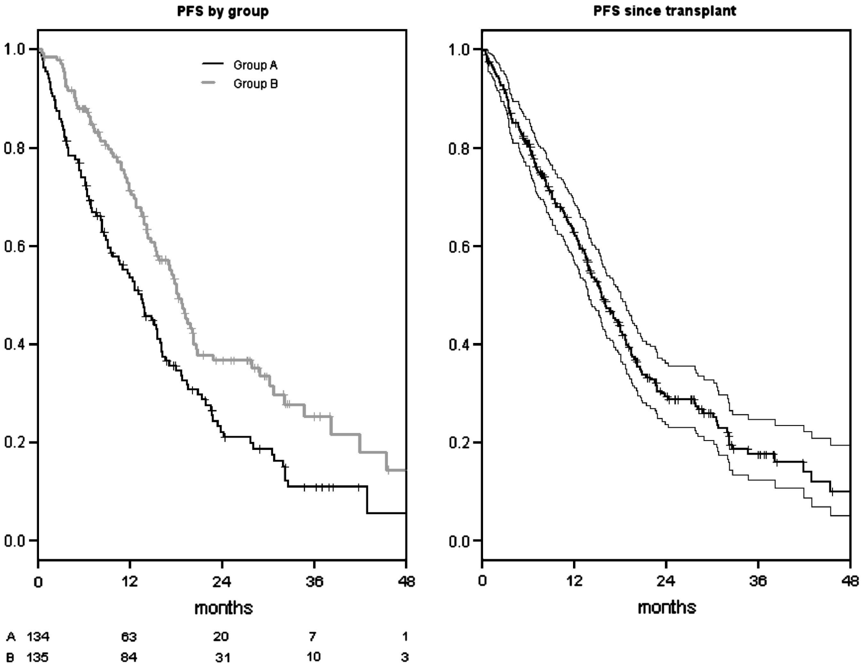

Figure 5 Kaplan–Meier curves. Left: a graph showing the number of patients at risk at relevant time points. Right: a graph showing the pointwise confidence bands.

reliability of a study are an 'honest' approach to the analysis that does not 'hide' unexpected results and does not force the model to prove what the investigator wishes to prove, the adoption of proper statistical methods, and a careful appraisal of the results that avoids over-interpretation and is performed within the framework of current clinical and biological knowledge.

4.2. Model building

4.2.1. The initial set of variables

The starting set of potential prognostic factors for the outcome of interest should of course include the main object of interest of the study plus all covariates known (by *a priori* knowledge, without looking at the current data) to have an influence on the outcome, especially those correlated with the main factor (as these are potential confounders). When the set of initial variables is too large, the statistician may begin excluding those variables that show the weakest evidence of a relationship with the outcome during marginal analysis; it is customary to consider a *P*-value threshold equal to 0.10 or 0.20, although it is a strong initial selection and not a recommended approach in general. In some cases with very large numbers of initial covariates, it may be useful to adopt some multivariate technique for data reduction, such as identifying two or three principal components from a large set of continuous variables and including them among the regressors, provided that they can be given a clinical interpretation.

In EBMT RBS, it is always advisable to include the calendar year, unless the population was chosen in a limited time span, because it captures sources of heterogeneity related to scientific progress and other changes in transplantation procedures that may have occurred during the study period (new drugs being used before or after transplantation, new diagnostic methods, admission to transplant of different types of patients, and so on). Care must be used when choosing the way in which to include the year. Treating it as a continuous covariate, in particular with a linear effect, might be incorrect,[30] while a meaningful categorisation might be appropriate. Additionally, the centre effect could be considered. Such an effect would represent both the potentially different standards of care (for example, more experienced centres may have a reduced incidence of transplant-related secondary negative events) or difference in procedures (preservation of cells, use of T-cell depletion); it would also represent the difference in patients in terms of characteristics that are either not observable or not recorded (for example, genetic differences; the EBMT registry currently includes centres from approximately 60 countries all over the world). Unfortunately, the centre effect cannot be considered a simple covariate unless the number of centres is very small (2–4 centres). Within a regression model, it can be included as a random coefficient, or the estimation of the effects and their standard deviations can be corrected to account for the dependence of observations from the same centre by using specific methods (see for example Andersen *et al.*,[31] Glidden[32] or Yamaguchi *et al.*[33]).

4.2.2. Identifying the shape of the effects

The inclusion of the effect of a covariate X through a term βX implies an assumption of linearity: any unity increase of X has the same effect (measured by β) on the outcome regardless of the starting value of X. This assumption may not hold in some cases; for example, a 5-year difference in age can have a different effect when comparing 30–35-year-old patients than when comparing 70–75-year-old patients. We can remove the linearity constraint by transforming X into $f(X)$ and including $\beta \cdot f(X)$ in the model. Identifying a correct shape $f(\cdot)$ of the effect of a continuous variable

X is important; in particular, a variable may appear non-significant only due to the fact that its effect strongly violates linearity (for example, it has a U-shaped effect, or an effect that is much stronger when X is low than when X is high). Possible simple transformations for $f(\cdot)$ are the logarithm or the power, but any mathematical function, even with a complex shape, can in principle be considered.

Sometimes the transformation of a continuous variable is suggested by substantive knowledge, or by its observed distribution. Thus, it is advisable to log-transform variables with a strongly right-skewed distribution (very extreme high values) because assuming linearity will likely overestimate the risk for very high values. Model validation techniques based on the analysis of residuals can suggest the correct shape for a transformation (see Section 4.3.4 for Cox regression). Some mathematical techniques allow the estimation of a flexible form of the effect of X from the data using functions such as splines or other combinations of polynomials.[18,28] The drawback of this latter approach is that the shape of $f(\cdot)$ is very well fit to the current data but may be not general enough when looking at other data. Moreover, the effect of X is nicely described graphically, but it is difficult to quantify numerically. Thus, the use of splines can be considered when an insight into the shape of the effect of X is an important target of the study; otherwise, choosing a simple transformation is more useful when the model is used to quantify the risk, or when it is necessary to generalise the model to other sets of data.

A special case of mathematical transformation for continuous variables is the categorisation into classes, using cut-points to define risk groups. Very often only two groups (1 cut-point) are chosen. Passing from a continuous to a categorical covariate is not recommended in view of the loss of information. Nonetheless, it is very common, and it is justified by the ease of describing the effect of X (for example, to produce survival curves). It is also usually performed when producing risk scores, although, as noted above, specific statistical techniques should be used for this purpose—techniques that are hardly seen in this type of study. In particular, producing models affected by overfitting and proposing data-driven scores that in practice will not perform well on different data sets should be avoided. More remarks against categorisation (in particular, dichotomisation) and additional references can be found in Royston *et al*.[34] One important 'minimal' requisite is that before transforming X into a dichotomous covariate, it should be verified that there is a dose–response effect for X, both as a continuous variable and as a categorical variable with several levels (see Section 4.2.3).

A good criterion to use when choosing cut-points is to base it on clinical/biological knowledge; in this case, the choice is not data-driven, and the results are comparable with the existing literature. However, sometimes it is not yet known how factor X affects the outcome, and it is a target of the study to achieve more knowledge or improve a prognostic scoring system based on the values of X. Thus, current data must sometimes be used to identify a categorisation. It is considered an 'objective' criterion (though it is data-driven) to create subgroups by cutting X at quantile values. For example, to create four subgroups of approximately the same size (with 25% of the patients in each), the cut-points would be Q1, the median and Q3. This approach may be inefficient, however, because it does not consider the effect of X on the outcome. In fact, other more or less refined 'outcome-oriented' methods exist. One example is a method proposed by Klein and Moeschberger[17] (par. 8.6); some statistical software programs also implement techniques of this type. A simple possibility is performing a residual analysis: if the shape $f(X)$ estimated from the residuals is a steep increase of risk around $X = c$, and before and after c the increase is rather small, then the choice could be to dichotomise X at c.

So far, we have considered the identification of the shape of the effect of a continuous variable X. The next section considers the effect of a categorical factor.

4.2.3. Including categorical covariates

Let us consider the case that X is a dichotomous covariate and assumes two possible values, 0 and 1, representing the absence and presence of a certain characteristic, respectively. The regression parameter β represents the effect of the presence of that certain characteristic compared to the case where it is absent.

If X has three (or, in general, k) possible levels coded as 0, 1, 2 (… up to k), then one of them is chosen as the reference level ('baseline') and the others are compared to the baseline. This is done either automatically by the software once it is instructed that X is a 'factor' or a 'categorical variable' (and the baseline is specified; by default, software programs may chose the first or the last level as the baseline), or by using indicator ('dummy') variables.

Table 5. Dummy variables for the phase of CML

	X	Dummy variables		
	Phase	CP	AP	BC
Chronic phase (CP)	0	1	0	0
Accelerated phase (AP)	1	0	1	0
Blast crisis (BC)	2	0	0	1

Table 5 shows the three dummy variables corresponding to $k = 3$ levels of the variable phase of CML. The effect of phase when assuming that chronic phase is the baseline is estimated by including in the regression model $(k-1) = 2$ dummy variables for the other two levels, that is, by including AP and BC among the covariates.

The regression coefficient β_{AP} (the beta for the variable AP) represents the effect of having CML at accelerated phase compared to chronic phase; similarly, β_{BC} (the beta for the dummy BC) represents the effect of having CML in blast crisis compared to chronic phase. The effect of BC versus AP is obtained as a difference: $\beta_{BC}-\beta_{AP}$. The effects for the dummy variables are tested separately, but the overall test for the presence of any effect of X (phase) should be performed by defining H_0 as $\beta_{AP} = \beta_{BC} = 0$ and H_1 as $(\beta_{AP} \neq 0)$ OR $(\beta_{BC} \neq 0)$ OR $(\beta_{AP} \neq \beta_{BC})$ and using the likelihood ratio test (or its alternatives, see note 38). Moreover, a correction for the inflation of type I error could be applied. The Bonferroni–Holm method is shown below.

The Bonferroni–Holm correction for multiple testing

The rule of Bonferroni–Holm is used to control for making a type I error in a situation with multiple testing (Section 1.4). This correction should be considered when testing the differences among more than two levels of a categorical variable, but we could also consider a situation where two groups are compared with respect to *k* different endpoints, or according to a predefined analysis plan where there was a set of hypotheses to be investigated; see Bauer[1] for references. The procedure is first described and then illustrated in Example 5, which assesses the effects of a categorical variable with three levels in a multivariate regression model.

There are *k* null hypotheses, each stating the absence of a difference (in general notation, H_{0j}: $\delta_j = 0$). We want to control the overall probability of making one or more false rejections (that is, rejecting a true null hypothesis) so that it is less than a pre-specified α (for example, $\alpha = 0.05$). We perform *k* tests, one for each null hypothesis, and then order the *P*-values in ascending order: $P_{(1)} \leqslant P_{(2)} \leqslant ... P_{(k)}$ For each value occupying the *j*-th place, there is a threshold with which to compare them, which is equal to $\alpha/(k-j+1)$. The comparisons are performed in sequence starting from $P_{(1)}$, and these comparisons are stopped when we have $P_{(j)} > \alpha/(k-j+1)$. All of the null hypotheses corresponding to the ordered *P*-values smaller than $p_{(j)}$ are rejected; the others cannot be rejected.

In the case that there are logical relationships between the hypotheses such that the number of true hypotheses can only be found in a set *X*, the Bonferroni–Holm procedure may be too conservative; in other words, it may encounter the opposite problem of failing to reject false null hypotheses. As a solution, the Shaffer correction suggests changing the thresholds for comparison by replacing the denominators with max{*x* in *X* such that $x \leqslant k-j+1$}.

Example 5: Testing the differences among the levels of a categorical factor (>2 levels)

Consider the effect of the factor phase of CML with levels chronic phase (CP, the baseline), accelerated phase (AP) and blast crisis (BC) in a multiple regression model. We have $k = 3$ tests for three comparisons: AP vs CP, BC vs CP, and BC vs AP. The corresponding null hypotheses are as follows: H_{01}: $\beta_{AP} = 0$ H_{02}: $\beta_{BC} = 0$ H_{03}: $\beta_{BC} - \beta_{AP} = 0$. We want to control the total 'false discovery' error probability so that it is lower that $\alpha = 0.05$. Performing a test for each null hypothesis, we obtain three *P*-values, say P_1, P_2 and P_3, and suppose that $P_2 < P_1 < P_3$.

- Compare P_2 with $\alpha/k = 0.05/3 = 0.017$. If $P_2 > 0.017$, then none of the hypotheses can be rejected and the phase of the disease is shown to have no impact.
- If $P_2 \leqslant 0.017$, then compare P_1 with $\alpha/(k-1) = 0.05/2 = 0.025$. If $P_1 > 0.025$, then only H_{02} can be rejected, and it is proven only that patients in blast crisis behave differently from the group in chronic phase.
- If $P_1 \leqslant 0.025$, then compare P_3 with $\alpha/(k-2) = 0.05/1 = 0.05$. If $P_3 > 0.05$, then H_{01} and H_{02} can be rejected, but not H_{03}, which means that the data do not support the hypothesis that blast crisis and accelerated phase behave differently. If $P_3 < 0.05$, then all three hypotheses can be rejected.

This is the Bonferroni–Holm correction. But, this is the situation considered by Shaffer, where only certain combinations of true/false for the three hypotheses make sense. In fact, it is not possible that only two of them are true and the third is not (for the transitive property, if $\beta_{AP} = 0$ and $\beta_{BC} = 0$, it necessarily follows that $\beta_{BC} - \beta_{AP} = 0$). Thus, the number of the true hypothesis can be 0, 1 or 3 and $X = \{0, 1, 3\}$.At this point, the threshold levels for the comparison of our *P*-values are:

- For the smallest *P*-value $P_{(1)}$ (P_2 in our example): the higher number in *X* that is $\leqslant k = 3$ is 3 $\alpha/3 = 0.017$.
- For the next *P*-value $P_{(2)}$ (P_1 in our example): the higher number in *X* that is $\leqslant k-1 = 2$ is $1\alpha/1 = 0.05$.
- For the highest *P*-value $P_{(k)}$ (P_3 in our example): the higher number in *X* that is $\leqslant k-2 = 1$ is $1\alpha/1 = 0.05$.

As can be seen, with this correction, it is easier to prove that H_{01} is false.

One kind of transformation of the effect of a categorical covariate with *k* levels is collapsing categories after having verified that the levels to be combined show no relevant different effect, and provided that the combined category has a sensible clinical interpretation.

A special type of transformation for ordered factors is assuming a linearity of the effect, or treating it as a continuous covariate. In our example, if phase is included as a continuous covariate, the effect of level 2 (BC) vs level 1 (AP) and the effect of level 1 (AP) vs level 0 (CP) are assumed to be equal and are measured by one single parameter β; the effect of level 2 (BC) vs level 0 (CP) will be measured to be equal to 2β. This assumption makes the model more parsimonious (one regression parameter is estimated instead of two) and the risk score easier to compute, but it must be verified. Fitting both the model with phase as a continuous variable and the model with the dummy variables AP and BC, the values β and 2β from the first model must be compared with β_{AP} and $\beta_{BC} - \beta_{AP}$, respectively, and with β_{BC} from the latter model. This type of investigation can also be considered when assessing the presence of a dose–response effect of a continuous variable *X* when looking for cut-points (Section 4.2.2).

4.2.4. The selection of variables

Generally speaking, the decision on whether a variable should be included in the model should take into account the clinical relevance of the variable (*a priori*), of its estimated effect (size and significance) and the interpretability of the model. Of course, the evaluation includes a careful investigation of the shape of the effect (Sections

4.2.2 and 4.2.3). In addition, the statistician should consider the consequences of the inclusion/exclusion in terms of overall significance of the model, stability of the effects of the other variables and validity of the model with respect to the assumptions (Section 4.2.6). In particular, if including or excluding a variable changes dramatically the effects of other covariates or makes more evident the violation of an important hypothesis (such as the proportionality of the hazard functions in the Cox regression), then it is necessary to understand why it is so by looking at the distribution of the variable, at its correlation with the other variables, or even at the influence of single observations.

Regarding how many variables a model can include, the rule of thumb is to limit the number so that there are 7–10 'units of information' (in survival analysis, only observed events, not censored observations, count as units containing information, while in logistic regression, it is the total sample size) per parameter to be estimated (a linear continuous variable effect requires 1, as with dichotomous variables; for a categorical variable with k levels you 'spend' $k-1$ degrees of freedom, and so on). Another golden rule of model building is parsimony. Here, the preferred model has only a few, relevant factors because it makes efficient use of the information for the estimation, it is simpler to interpret and it is more useful for being generalised to populations other than the source population. Too many prognostic factors may in fact lead to overfitting, where the model fits the current data very well, but is too 'tailored' to the data, and therefore, it may be inadequate for other data sets. For the purpose of controlling the number of covariates (more precisely, of parameters) being included in a regression model, some texts suggest rules based on the Akaike information criterion.[39]

Identifying a model may actually require a circular process. The effect of a variable or its shape can change depending on the inclusion or exclusion of other variables, and if a model appears unsatisfactory at the validation step, it should be better identified. The complexity of the process and of the issues involved is the main reason to avoid the automatic selection procedures that are implemented by many statistical software programs (sometimes called 'stepwise' regression routines[40]). Given that they all rely only on significance without taking into account relevance, interpretation, or any other 'good sense' or methodological issue, automatic processes should be avoided, or used only to select among variables with similar meaning, or when verifying the presence of interactions.

4.2.5. Interactions

A statistical interaction between two variables X_1 and X_2 is a modification of the effect (on a certain outcome variable Y), and it occurs when the effect of one changes depending on the level of the other.[41] A statistical interaction can also have a clinical interpretation; for example, a risk factor X_1 has less impact in younger patients than it does in the elderly (X_2 is age), or the gender mismatch effect in SCT (male recipients with female donors tend to have a higher risk of death in many diseases) can be seen as an interaction between patient gender and donor gender.

From a practical point of view, the interaction term is usually introduced in the model containing X_1 and X_2 as their product $X_1{*}X_2$ along with its beta:[42]

$$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$$

Then, for two dichotomous covariates the effect of $X_1=1$ compared to $X_1=0$ is equal to β_1 when $X_2=0$ and it is equal to $\beta_1+\beta_3$ when $X_2=1$. When X_2 is continuous, the effect of X_1 depends on the value of X_2, being equal to $\beta_1+\beta_3 \cdot X_2$. The test for H_0: $\beta_3=0$ returns an assessment of the significance of the interaction between X_1 and X_2.

The investigation of the presence of certain interaction terms can be a specific objective in a study as a valid approach alternative to subgroup analyses because there is more efficient use of the information and the possibility of testing the difference of an effect among subgroups.

Verifying the presence of interaction terms is also a final phase of model identification. In the absence of clinical criteria, all pairwise interactions between all the variables included could, in principle, be assessed, but this would lead to problems such as overfitting or multiple testing (even if no interaction is present in the population, some terms may turn out to be 'significant' by pure chance). In practice, especially if there are many adjustment covariates and there is a main factor of interest, the assessment could be restricted to the interactions of the main factor with all of the other covariates. In addition, significance has little importance with respect to two other aspects: whether the effect modification is relevant in size, and whether it is interpretable in clinical or biological terms. A significant interaction term that produces a rather irrelevant effect modification may be neglected for the sake of parsimony and simplicity of the model, while even a weakly significant interaction (for example, a P-value below 0.1 or 0.2, but not below 0.05) for which there exists a clinical interpretation could be kept in the model if it indicates a strong change of the effects. An uninterpretable (significant and relevant) interaction could be an interesting result to investigate further, but it could also be found due to pure chance or due to a model misspecification.

These comments suggest defining in advance a few interactions to examine on the basis of the literature and

[39] $\text{AIC}=-2\cdot\text{Log-likelihood}+kp$, where p is the number of (relevant) parameters in the model, and k is a constant, usually 2. This quantity decreases as p begins to increase, up to the point where unnecessary variables are included. For its application in the framework of the Cox regression, see for example Klein and Moeschberger,[17], par. 8.7.

[40] The 'forward' approach adds the variables one after another, choosing the most significant at each step, until no variable adds significant information. The 'backward' procedure starts with all variables and removes the least significant variable at each step, until the loss of information becomes significant (between these two, the latter approach is preferred, unless the initial set is very large). Some software programs also implement combinations of forward and backward methods, or other methods (for example, 'best subset').

[41] Interactions among more than two variables (typically three) can in principle be considered, but they hardly correspond to real biological phenomena. They are in any case very difficult to interpret, and thus are rarely seen.

[42] To assess the presence of an interaction term $X_1{*}X_2$, the model must be 'hierarchical', that is, it must include both main effects X_1 and X_2.

other considerations based on the current knowledge of the phenomena being investigated, and to avoid (as usual) decisions for model building based solely on significance.

As a final suggestion, when a certain adjustment variable expected to be influential for the outcome appears negligible according to the selected model, it may be worth assessing if it acts on the outcome only through an interaction with another covariate.

4.2.6. Model validation/diagnostic

This section focuses on methods for internal validation, or diagnostics, the assessment of the validity of the fitted model under many respects, including substantial validity of the assumptions on which the model is based in the observed data, goodness of fit (that is, good estimation of the outcome from the explanatory variables) and robustness with respect to particular observations. Several validation methods exist depending on the type of regression model used and on which aspect of the model is being verified. Many methods are based on the analysis of residuals, quantities that (generally speaking) express the distance between the observed data and the predictions obtained from the model. Others are based on measures of the impact of each observation on the estimated model (influence analysis).

Despite introducing this phase of model building at the end of the section on regression, the model can be validated at any step because validation also indicates how to correct the model specification. For example, in Cox regression, a certain method of residual analysis can be used to check the validity of the hypothesis of proportionality of hazards for the covariates included in the model, and can also indicate how to correct for non-proportionality. We also saw the use of validation techniques based on residuals to identify the functional form of the effect of a continuous covariate (Section 4.2.2).

The amount of refinement of model building depends on several aspects, but generally speaking the statistician should be able to propose a model substantially valid with respect to the main assumptions, and stable with respect to small variations of the sample (for example, not influenced by single, extreme observations) or to the inclusion/exclusion of the variables. Increasing the goodness of fit should not be considered a main objective of model building because there are several negative consequences in case of overfitting: the prediction performance of the model will be poor (that is, the model may fit poorly with the outcomes for another sample of data), and the prognostic value of a factor or of derived risk scores will be small on an independent data set.

4.3. The Cox model

The Cox model is currently the most widely used regression model for survival data.[43] It is characterised by two features: the assumption of PH and the lack of specification of a functional form for the dependence of the hazard function over time. The PH assumption was introduced in Section 2.2. In the Cox model, it works by assuming that the covariates act multiplicatively on the hazard function, which depends on time only through a baseline hazard function h_0:

$$h(t; x_1, x_2, \ldots, x_k) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

The effect of a unit increase of X_2 on the hazard while keeping all other covariates X_j constant is measured by the HR and obtained by exponentiating the coefficient β_2:

$$\frac{h(t; x_1, x_2 + 1, \ldots, x_k)}{h(t; x_1, x_2, \ldots, x_k)} = \exp(\beta_2)$$

It is worth recalling the important interpretation that the HR is constant in time; that is, an increase of X_2 has the same effect on the hazard at the start of the follow-up as at any time, even in the very long term. There is no change of the effect over time. The effect of time is included only in the baseline hazard. In the Cox model, the use of an estimation technique based on a partial likelihood allows us to leave the functional form of the baseline hazard unspecified and thus to neglect it. This regression technique focuses on the estimation of the effects of covariates in relative terms; note in particular that the Cox partial likelihood method does not produce an estimate of the hazard corresponding to a certain covariate pattern and at a certain time t. The latter can be obtained using the betas estimated with Cox together with a non-parametric estimate of the cumulative baseline hazard; this is useful when plotting survival curves to represent graphically the effect of covariates.

4.3.1. HRs in practice

- For a continuous covariate X with an estimated coefficient β and $\theta = \exp(\beta)$, the HR measuring the effect of an increase of, for example, 5 units is obtained as θ^5.
- For a categorical covariate with three levels (or in general, k levels), with the first being the baseline and the other two being represented by two dummy variables, the effect of each level vs the baseline is given by HR $\theta_j = \exp(\beta_j)$, and the HR between the two is obtained by the ratio θ_{j1}/θ_{j2}. In the example of the phase of CML in Section 4.2.3, the HR to compare BC to AP is $\theta_{\mathrm{BC}}/\theta_{\mathrm{AP}}$.
- If a three-level ordered categorical covariate is included as a numeric variable with linear effect (Section 4.2.3), then the HR of level 1 vs level 0 and of level 2 vs level 1 is the same and is equal to $\theta = \exp(\beta)$; the HR of level 2 vs level 0 is θ^2.

4.3.2. The stratified Cox model

A stratified Cox model has a different baseline hazard for each of k subgroups of patients defined by a k-level categorical variable (stratification factor) and assumes that the effects of the other covariates are the same in each stratum:

$$h(t; x_1, \ldots, x_k, \mathrm{stratum}\, j) = h_{0j}(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

The assumption that all other covariates act in the same way in each stratum can be relaxed by including stratum-

[43] Because it is a model of a hazard function, it can also be used to model the cause-specific hazard of an event with competing risks (Section 2.2.2).

specific covariates. For example, with two strata, 1 and 2, and X with different effects, the model will include:

$$X_1 = \begin{cases} X \text{ in stratum } 1 \\ 0 \text{ else} \end{cases} \text{and } X_2 = \begin{cases} X \text{ in stratum } 2 \\ 0 \text{ else} \end{cases}$$

If a categorical variable X has an impact on the outcome that cannot be assumed to be constant in time (that is, it violates the PH hypothesis), then stratifying the model on X removes the problem of violating proportionality, with the drawback being that the effect of X is only ‘removed’, but not estimated from the model.

4.3.3. Time-dependent covariates

Time-dependent covariates are, generally speaking, variables whose value changes with time. They are used in three situations:

- When a characteristic changes during follow-up, such as when a treatment is changed according to protocol or the level of a certain biological parameter monitored during follow-up changes.
- More specifically, when there is a change in the status of the patient, which is relevant for subsequent outcome; for example, when an event such as GVHD or second transplant occurs (Section 2.1.6).
- When a covariate shows a non-proportional effect (see Sections 4.3.4 and 4.3.5). In fact, a time-varying effect of X would require a non-constant beta and thus $\beta(t) \cdot X$, which can be represented by $\beta \cdot X(t)$. In other words, to represent a time-varying effect for a constant covariate, we can include a constant effect for a time-varying covariate.

The variation of a covariate in time is addressed by splitting the individual follow-up time into time periods where the value of the covariate is constant, and replacing the event history of a patient with a series of histories along consecutive periods; left-truncation (Section 2.1.4) is used to account for delayed entry. An example is given in Table 6 for the case that X represents the occurrence of a second transplant. The procedure is valid for any case of time-varying covariate; when it varies continuously, the follow-up is split into very small intervals.

Notice that in practice the majority of statistical software programs implement time-dependent covariates within the routines for Cox regression through specific programming instructions. Thus, the user does not truly have to manipulate the data set directly.[44]

Example 6: Use of a time-dependent covariate to include the effect of second transplant in a Cox model

The computation of the time-dependent covariate is illustrated here for two observations. Patient A does not receive a second SCT and dies at time 4 while Patient B receives a second SCT at time 2, and then dies at time 7. The latter event history is split into two periods. In the first period from time 0 to time 2, X is constantly equal to zero, the patient has had no second SCT (yet), and the final status is always 0, censored. The second period extends from time 2 (delayed entry) to time 7, and X is equal to 1, a second transplant was given, and the final status is equal to the original one (in this case, the patient died). All other covariates for B are the same in the two rows. For patient A, there is only one row for the time period from 0 to 4, and the survival status is as observed. The expanded data set will thus have three rows to represent these two patients.

Table 6 Follow-up split for a time-dependent covariate X

Patient id	Gender (M = 1, F = 2)	Tstart	Tstop	X (2nd SCT); No = 0, Yes = 1	Surv (survival status); no = alive = 0, yes = dead = 1
A	2	0	4	0	1
B	1	0	2	0	0
B	1	2	7	1	1

The outcome variable in the Cox model will be the triplet given by the starting time, the end time and the survival status, which accounts for delayed entry (Tstart, Tstop, Surv). The HR for the covariate X compares the hazard of death of two patients, alive at the same time, with the same characteristics, except that one received a second transplant before the time of comparison and the other did not. For the interpretation of this HR, see the remarks in Section 2.1.6.

4.3.4. Validation of the Cox model

There are several methods for validating the Cox model as illustrated by Klein and Moeschberger[17] (Chapter 11), Therneau and Grambsch[18] (Chapters 4–7), and Hosmer and Lemeshow[19] (Chapter 6). It is worth suggesting a careful use of these techniques, especially when building a model for prediction purposes.[28] This section briefly introduces two validation methods based on the analysis of residuals that are related to important aspects of model building:

- the use of martingale residuals for the identification of the shape of the effect of a continuous covariate (Section 4.2.2); and
- the use of (scaled) Schoenfeld residuals for verifying the PH assumption.

Both techniques include a graphical evaluation based on a scatter plot of residuals interpolated with a smooth line to highlight trends of departure from the null, which is the reference value corresponding to the ‘perfect’ fit.

Each model fit returns a set of martingale residuals, one for each observation included in the sample. The residual represents the difference between the actual survival status observed and the risk of failure predicted by the model; a positive (negative) residual indicates that the risk is underestimated (overestimated). In an ideal situation of perfect fit, the residuals should be equal to zero; with a properly specified model, they should vary randomly around zero. If the model is not well specified with respect to the effect of a variable X, then the residuals tend to

[44] In R this manipulation is necessary. Some libraries provide functions to do it, for example `mstate` or `Epi`. Continuous $X(t)$ can be computed using the command survSplit from the in-built `survival` library.

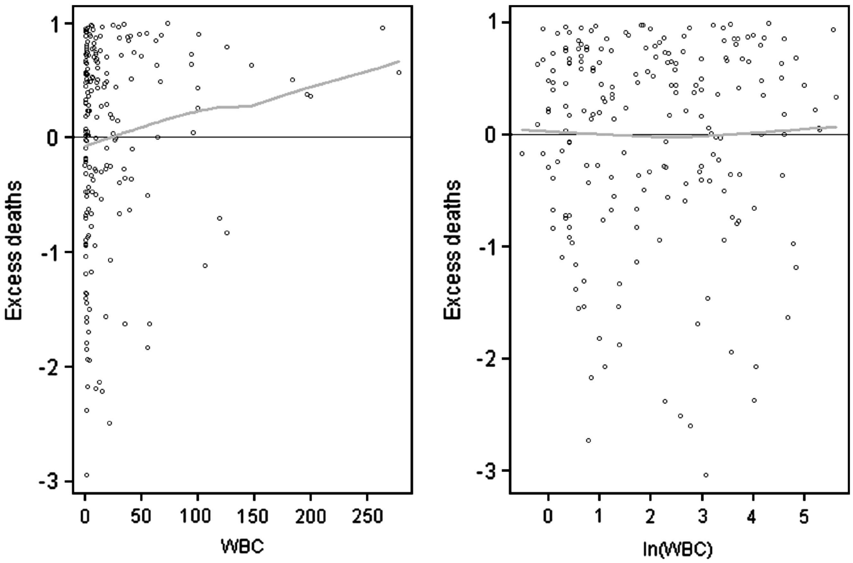

Figure 6 Illustration of the use of martingale residuals to identify the effect of a covariate X in the Cox model. Left: residuals from a model not including X. The graph suggests including X as a risk factor: in fact, the estimated risk is too high for small values of X and too low for large values of X. Linearity appears to be inappropriate, as the slope is not constant. Ln(X) appears to be a proper choice. Right: residuals from a model including ln(X). The residuals distribute approximately at random around 0 without apparent trends related to the covariate ln(X): the shape is appropriate.

distribute according to a specific trend dependent on the value of X and not randomly around zero. This allows a graphical evaluation of the shape of the effect of X, as illustrated in Figure 6.

At each model fit, a series of (scaled) Schoenfeld residuals for each observed failure is estimated for each covariate included in the model. If the effect of a covariate X is constant over time, these residuals distribute at random around the reference value zero, without trends associated with the observed failure times. If, however, the effect of X depends on time, such that we can think of a regression coefficient beta that is a function of time, $\beta(t)$, then a graph of the residuals vs the observed failure times (or a transformation of them, such as ranks) will suggest the shape of $\beta(t)$, as shown in Figure 7. The observed trend can be tested for significance; for example, Therneau has developed a routine in R (`cox.zph`) for performing the tests for the (linear) trends for each covariate and for performing a global test for the hypothesis that PH holds.

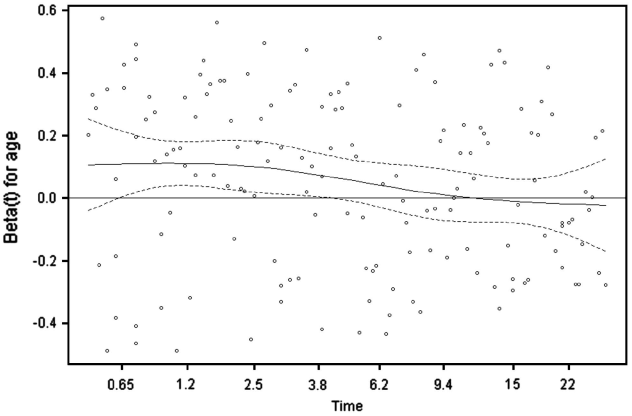

Figure 7 Use of scaled Schoenfeld residuals to detect non-PH for the effect of a covariate 'Age' in the Cox model. The residuals from the model including X are plotted in this case vs the ranks of the observed failure times. The graph suggests that the covariate ('Age') has a decreasing effect as risk factor, which tends to become nonsignificant in time (the confidence interval includes the straight line through 0).

Another customary way to check for PH is by including some time-varying effect of X through a time-dependent variable representing the interaction of X with time, such as $X \cdot \log(t)$ or Xt (Section 4.3.3), and testing its significance. This method has the limitation that the assessment is performed only for a specific type of violation of the independence on time.

4.3.5. When the PH assumption does not hold

Non-PH for a covariate X can be explained in terms of auto-selection of the 'fittest' patients in the subgroup with higher risk, for which the difference from the low-risk patients fades over time. This mechanism is formally illustrated in the context of frailty models[18,35] that consider the presence of unobserved heterogeneity, a source of variation that is not imputable to the covariates included in the model.[45] In some cases, there is an expected biological explanation, such as when a characteristic is a risk factor for early post transplant infection, and thus, this mechanism has an effect on the risk of death that diminishes and then disappears over time. In another situation, the unobserved characteristic may have an effect when it is

[45] Frailty models are extensions of the Cox regression to include random effects, represent unobserved heterogeneity or create dependence among observations from the same cluster, such as multiple endpoints for the same patient, or more simply observations from patients grouped according to the centre (to account for the centre effect). The latter situation can also be addressed with a 'marginal' approach that corrects the variance and covariance matrix (in R, see the options 'frailty' and 'cluster' within the coxph procedure, respectively).

associated with a protection against relapse, and then the hazards will tend to diverge in the long term. A strict proportionality of hazards is perhaps not met in the majority of cases, but strong violations of the PH assumption affect the validity of the Cox model. When the effect of X is non-proportional, the true HR is time-dependent, and it should be described by a function HR(t); the estimated time-fixed HR is some kind of weighted average of HR(t), and the corresponding significance test does not refer to the effect at a specific point in time. The Cox model can be amended for non-proportionality of the effect of X in two ways:

- by including X as a stratification factor; or
- by adding a time-varying effect for X using a time-dependent covariate.

Stratification is a good solution when the X is categorical or it can be categorised to define clinically relevant subgroups and its effect is not an object of interest.

Alternatively, when the time-varying effect of X is of interest, the model must include X plus a time-dependent covariate $X(t) = X \cdot f(t)$ (as in fact $X(t)$ represents the interaction of X with a function of time, $f(t)$) (Section 4.3.3). Typical choices are $f(t) = t$ or $f(t) = \log(t)$ to represent effects that are at first rather stable but then disappear after some time, or step functions. A common approach to gaining some insight into early and late effect is by including two covariates:

$$X_1 = \begin{cases} 1 & \text{if time} \leq \tau \\ 0 & \text{else} \end{cases} \quad \text{for early effect,}$$

$$X_2 = \begin{cases} 1 & \text{if time} > \tau \\ 0 & \text{else} \end{cases} \quad \text{for late effect}$$

As was seen in Section 4.3.4, the function $f(t)$ may be identified by analysing the Schoenfeld residuals. The computation of $X(t) = X \cdot f(t)$ is performed by splitting the follow-up and expanding the data set as illustrated in Section 4.3.3 (beware of this common mistake: a simple multiplication of the columns X and $f(T)$, where T is the observed survival time, is NOT a time-varying covariate). With the model

$$h(t; x, (\text{other covariates})) = h_0(t) \exp(\beta_1 x + \beta_2 x \cdot f(t) + \ldots)$$

the HR for a unit increase of X depends on the time t and is computed as:

$$HR(t) = \exp(\beta_1 + \beta_2 \cdot f(t))$$

In extreme situations of departures from PH, such as when the hazard functions cross, the results of a Cox regression may be completely misleading. When a covariate X defines two hazard curves that cross, despite a substantial difference in long-term survival, because of early differences in the opposite direction, the estimated HR from a Cox model may be close to zero. Even if the model is amended by adding time-dependent covariates for the effect of X, the hazard function HR(t) is rarely a true object of interest, and in particular a global test for HR(t) is rather useless. The primary research questions would normally focus either on short- or long-term outcomes, and there are specific methods to address each case.

The long-term outcomes are often of primary research interest. When focusing on the analysis of the survival probability at specific points in time, methods such as those proposed by Klein *et al.*[13] and Logan *et al.*[14] should be used. In particular, approaches based on pseudo-values also allow adjusted comparisons. Another issue of interest is the probability of cure, or the probability that the risk of failure decreases towards levels that are comparable to those of the general population. In paediatric oncology research, the fraction of cured patients is particularly relevant. Cure models[15,16] investigate cure rates and survival times separately. In the presence of a cure and when follow-up is sufficient, they are more efficacious in detecting treatment differences with non-PH than a Cox regression.

Finally, another possible approach to the analysis of time-varying effects is to apply models that do not require PH. There are several alternative methods to the Cox regression, although they are (still) seldom used in clinical applications. For example, the additive model by Aalen, as well as some generalisations of it,[36] and the method of dynamic prediction by landmarking[37] are two possibilities. See also the 'Inventory' by Latouche[26] for more recent methods of survival analysis that serve as alternatives to the 'classic' methods.

4.4. Stratification, matching and propensity scores°

The methods of traditional epidemiology used for controlling confounding, such as stratified comparisons and matched-paired analysis (Section 1.3), are inferior with respect to regression modelling because the latter is capable of using all of the information available, controlling several confounders simultaneously, and providing correct and precise estimates of the effects/differences of interest. However, those methods become more interesting when stratification and matching are based on propensity scores (PSs), instead of on one or two covariates as in the traditional approach.

The framework where PSs are used is the comparison of two treatment groups (say, treated and not-treated) in a non-randomised study.[46] The main problem is that the two groups differ with respect to many characteristics, which actually could have determined the decision to treat or not. Thus, the idea is to compare each treated patient to a not-treated patient who, on the basis of the characteristics, had the same probability of getting the treatment; this probability is called the PS.[47]

[46] This is also the context where adjusted survival curves[42] can be used, see note 8.

[47] This idea is theoretically supported by work initiated by Rosenbaum and Rubin,[46] which shows that: (1) given the PS, the distribution of the covariates is the same in treated and not-treated groups, (2) under a condition called 'strong ignorability' (which corresponds to the assumption that there is no unmeasured confounder and that given the covariates there is no certainty regarding which treatment the patient will receive) and given a fixed PS, the difference of outcomes in treated and not-treated groups yields an unbiased estimate of the treatment effect.

The PSs are computed for each patient from the covariates, usually applying logistic regression; some statisticians criticise the habit of including a very large number of covariates in this model for PS, and recommend principally using covariates with some effect on the outcome being compared. Then, the analysis of the treatment effect can be performed in different ways.

One approach is the matched-pair analysis, possibly adjusted for other covariates (those associated with the outcome). For each treated patient, one or more 'control' not-treated patients are chosen from among those with the 'same' PS (for proper selection methods, see D'Agostino[38]). It must be verified that the two groups selected for the analysis are similar with respect to the characteristics (possibly avoiding reliance on significance tests; standardised differences may be used instead[48]). As in any matched-pair analysis, it is fundamental that the statistical methods used to compare treatments account for the dependence within each pair. (Thus, it is recommended to use paired *t*-tests, McNemar's test, conditional logistic regression, stratified Cox regression, and so on.) As in any 'traditional' matched-pair analysis, the drawback is a loss of information because of the reduced size of the control group. However, this approach is a good solution when there are few patients in the 'study' group, where a regression model would allow control of only a few covariates.

Another approach is the stratified estimation of treatment effect. Observations are stratified in a few groups (usually 5, at most 10) on the basis of quantiles of the distribution of PS and then estimates of the effect of treatment in each stratum are computed and combined using some weighted average. We will not go any further in the illustration of this approach, which is rather unsatisfactory,[39,40] and should in any case be improved against the risk of residual confounding by obtaining the effect estimate in each stratum through regression, including covariates.[39]

Other estimators of the treatment effect are derived using the inverse-weight technique. It was shown[39] that including quantities derived from regression in the estimator yields an efficient estimator that is also robust with respect to model misspecification. However, the illustration of these methods is beyond the scope of this document.

Although these methods may appear unduly complicated, many statisticians agree that they have better properties than a simple regression model adjusted for PS, treatment and perhaps other covariates.[38–40]At this point, one could wonder what is gained by using PSs instead of applying a regression on all relevant covariates as illustrated in Section 4.1. It must be said that among statisticians, there is no consensus on this. One argument in favour of PS is that, when the two groups are indeed quite different with distributions of covariates of X overlapping only in certain regions, the relationship between covariates and response is determined only by the treated patients in one region, and only by the not-treated patients in another, which means that in fact the relationship is extrapolated. However, if the two groups are strongly incomparable with respect to relevant characteristics, are we sure they should be compared at all? Perhaps the statistician and the investigators should consider whether it would make sense to compare only cases with similar distributions while leaving out the most extreme cases; then, the usual regression could be sufficiently useful.

Conflict of interest

The author declares no conflict of interest.

Acknowledgements

L. de Wreede contributed to the final version. All members of the EBMT Statistical Committee reviewed the Guidelines and agreed with the final version. SI has received grant support through her research position at the Tor Vergata University of Rome.

References

1 Bauer P. Multiple testing in clinical trials. *Stat Med* 1991; **10**: 871–890.

2 Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley: New York, 2004.

3 Beyersmann J, Gastmeier P, Wolkewitz M, Schumacher M. An easy mathematical proof showed that time-dependent bias inevitably leads to biased effect estimation. *J Clin Epidemiol* 2008; **61**: 1216–1221.

4 Simon R, Makuch RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. *Stat Med* 1984; **3**: 35–44.

5 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007; **26**: 2389–2430.

6 Keiding N, Klein JP, Horowitz MM. Multi-state models and outcome prediction in bone marrow transplantation. *Stat Med* 2001; **20**: 1871–1885.

7 Iacobelli S, Apperley J, Morris C. Assessment of the role of timing of second transplantation in multiple myeloma by multistate modeling. *Exp Hematol* 2008, 1567–1571.

8 Iacobelli S. Statistical modeling of complex disease histories in Bone Marrow Transplant. Guidelines for proper use and interpretation of the Cox model for the European Group for Blood and Marrow Transplantation. 2004. Available from the EBMT website www.ebmt.org.

9 van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Anal* 2008; **14**: 447–463.

10 Klein JP, Rizzo JD, Zhang MJ, Keiding N. Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part 1: Unadjusted analysis. *Bone Marrow Transplant* 2001; **28**: 909–915.

11 Klein JP, Keiding N, Shu YY, Szydlo RM, Goldman JM. Summary curves for patients transplanted for chronic myeloid leukaemia salvaged by a donor lymphocyte infusion: the current leukaemia-free survival curve. *Br J Haematol* 2000; **109**: 148–152.

12 Liu, Logan, Klein JP. Inference for current leukemia free survival. *Lifetime Data Anal* 2008; **14**: 432–446.

13 Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. *Stat Med* 2007; **26**: 4505–4519.

[48] For example, the standardised difference of two averages is: $\frac{100\cdot|\bar{x}_1-\bar{x}_0|}{\sqrt{\left(s_1^2+s_0^2\right)/2}}$, where s_j^2 for $j=0$, 1 is the variance in group j.

14 Logan B, Klein JP, Zhang MJ. Comparing treatments in the presence of crossing survival curves: an application to Bone Marrow Transplantation. *Biometrics* 2008; **64**: 733–740.
15 Corbiere F, Joly P. A SAS macro for parametric and semiparametric mixture cure models. *Comput Meth Programs Biomed* 2007; **85**: 173–180.
16 Sposto R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Stat Med* 2002; **21**: 293–312.
17 Klein JP, Moeschberger ML. *Survival Analysis. Techniques for Censored and Truncated Data*. 2nd edn. Springer: New York, 2003.
18 Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer: Berlin, 2000.
19 Hosmer D, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd edn. Wiley: New York, 2008.
20 Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representation of old estimators. *Stat Med* 1999; **18**: 695–706.
21 Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat* 1988; **16**: 1141–1154.
22 Fine JP, Gray RJ. A proportional hazard model for the subdistribution of a competing risk. *JASA* 1999; **94**: 496–509.
23 Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudo-values of the cumulative incidence function. *Biometrics* 2005; **61**: 223–229.
24 Dignam JJ, Kocherginsky MN. Choice and interpretation of statistical tests used when competing risks are present. *J Clin Oncol* 2008; **26**: 4027–4034.
25 Logan BR, Zhang MJ, Klein JP. Regression models for hazard rates versus cumulative incidence probabilities in haematopoietic cell transplantation data. *Biol Bone Marrow Transplant* 2006; **12** (Suppl 1): 107–112.
26 Latouche A.. Improving statistical analysis of prospective clinical trials in stem cell transplantation. An inventory of new approaches in survival analysis'. *Technical Report of the CLINT—Establishment of infrastructure to support International Prospective Clinical Trials in Stem Cell Transplantation*, 2010. Available from COBRA Preprint Series, Art. 70. http://biostats.bepress.com/cobra/ps/art70 .
27 Klein JP. Modeling competing risks in cancer studies. *Stat Med* 2006; **25**: 1015–1034.
28 Harrell Jr FE. *Regression Modeling Strategies*. Springer: Berlin, 2001.
29 Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Control Clin Trials* 1996; **17**: 343–346.
30 Klein JP, Rizzo JD, Zhang MJ, Keiding N. Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part 2: Regression modelling. *Bone Marrow Transplantation* 2001; **28**: 1001–1011.
31 Andersen PK, Klein JP, Zhang MJ. Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Stat Med* 1999; **18**: 1489–1500.
32 Glidden DV, Vittingho E. Modeling clustered survival data from multicentre clinical trials. *Stat Med* 2004; **23**: 369–388.
33 Yamaguchi T, Ohashi Y, Matsuyama Y. Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Stat Meth Med Res* 2002; **11**: 221–236.
34 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006; **25**: 127–141.
35 Statistical Methods in Medical Research 1994 Vol. 3 (Five papers on frailty models for heterogeneity and dependence).
36 Scheike TH, Zhang MJ. Extensions and applications of the Cox-Aalen survival model. *Biometrics* 2003; **59**: 1036–1045.
37 van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scand J Stat* 2007; **34**: 70–85.
38 D'Agostino Jr RB. Tutorial in biostatistics. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
39 Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; **23**: 2937–2960.
40 Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Stat Med* 2007; **26**: 5529–5544.
41 Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. *Stat Med* 1997; **16**: 833–839.
42 Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Meth Prog Biomed* 2004; **75**: 45–49.
43 Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd edn. Wiley: New York, 2002.
44 Fine JP, Jiang H, Chappell R. On semi-competing risks data. *Biometrika.* 2001; **88**: 907–919.
45 Scheike TH, Zhang MJ. Flexible competing risks regression modeling and goodness-of-fit. *Lifetime Data Anal* 2008; **14**: 464–483.
46 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effect. *Biometrika* 1983; **70**: 41–55.
47 Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. *Bone Marrow Transplant* 2007; **40**: 381–387.
48 Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: an in depth guide for clinicians. *Bone Marrow Transplant* 2010; **45**: 1388–1395.
49 de Wreede L, Fiocco M, Putter H. mstate. An R package for the analysis of competing risks and multi-state models. *J Stat Softw* 2011; **38**: 1–30.
50 de Wreede L, Fiocco M, Putter H. The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. *Comput Meth Prog Biomed* 2010; **99**: 261–274.

Appendix

Procedures/commands in R, SAS and SPSS

For the purpose of helping new users perform their analyses, we provide here a list of procedures and instructions for the methods indicated in the previous chapters that are available in the statistical software packages R, SAS and SPSS. These three software programs are very commonly used in clinical analysis, but other reliable programs may provide good tools for survival and event-history analysis, for example STATA.

R is a reliable software package that is distributed freely on the web (http://www.r-project.org).[49] Although it may appear difficult to use, we encourage investing some time and patience in learning how to use it, especially if the user is willing to apply more than the usual standard methods of survival analysis. Consider also that the majority of basic statistical procedures can be applied through a menu-based interface (R Commander, for which you will need to install the library `Rcmdr`), and that there is abundant material available to facilitate the use of this program.

This list is not an exhaustive companion for statistical analysis; it is only intended to help new users quickly find material to start their search in the Help pages of the software.

[49] The website provides the R software, general manuals, and other useful material or links. Additional programs for specific methods are available from local websites (R Cran). The programs (all documented with help files and a manual in pdf format) can be easily and quickly installed online or downloaded as archive files and then installed.

Table 7: Procedures in R, SAS and SPSS for the description of categorical and continuous variables

	R	SAS	SPSS
Description			
Categorical variables → tables	Table	proc freq	frequencies; crosstabs; tables
Quantitative variables → indexes (and graphs)	summary; quantile (boxplot; hist)	proc univariate; proc means	frequencies/for = not/stat = min max mean median; summarize; examine; descriptives; means
Differences in k groups of a continuous, normal variable			
$k = 2$, *t*-Test	t.test	proc ttest	t-test
$k > 2$, ANOVA	aov, lm, anova	proc anova	anova; oneway; summarize; glm
Non-parametric tests for differences in k groups			
Continuous variable, $k = 2$, Mann–Whitney test	wilcox.test	proc npar1way	npar tests
Continuous variable, $k > 2$, Kruskal–Wallis	kruskal.test	proc npar1way	npar tests
Categorical variable, χ^2-test	chisq.test	proc freq/chisq	crosstabs /stat = chisq; npar tests
Categorical variable, $k = 2$, Fisher exact test	fisher.test	proc freq/exact	crosstabs /stat = sher; npar tests
Association between two continuous covariates			
Linear correlation	cor(method = ‘‘pearson’’)	proc corr pearson	correlation
Association in general, non-parametric tests	cor(method = ‘‘kendall’’, ‘‘spearman’’)	proc corr kendall spearman	nonpar corr

Table 8: Procedures in R, SAS and SPSS for the description of the occurrence of events

	R	SAS	SPSS
Survival-like endpoints	library(survival)		
Kaplan–Meier estimates	survt	proc lifetest	km
Log-rank test and others	survdiff	proc lifetest	option /compare in km
Cox regression	coxph	proc phreg	Coxreg
Competing risks	library(cmprsk)	macros exists	macros exists
Cumulative incidence estimates	cuminc		-
Gray test	cuminc		—
Fine and Gray regression	crr		—
Events before a certain time, complete follow-up			
Regression models	glm (family: binomial, link = ’’logit’’)	proc logistic; proc catmod	logistic regression; genlog

The R library cmprsk has to be downloaded from a R Cran and installed; references for using it are the manual of the library, by Gray and the papers (proposing complementary software) by Scrucca *et al.*[47,48] Also the library mstate (created for multi-state models) can compute cumulative incidence curves. See the papers by de Wreede *et al.*[49,50] Many useful SAS macros, including for competing risks, can be found from the Wisconsin Medical College website. Most of the procedures described in these guidelines are readily available in SPSS, an important exception being the methods for competing risks; an SPSS macro for the cumulative incidence estimator was created by S Le Cessie of the Leiden Department of Medical Statistics, and is available from her website.