Universiteit Leiden
The Netherlands

Analysis of survival outcomes in haematopoietic cell transplant studies: pitfalls and solutions

Wreede, L.C. de; Schetelig, J.; Putter, H.

Citation

Wreede, L. C. de, Schetelig, J., & Putter, H. (2022). Analysis of survival outcomes in haematopoietic cell transplant studies: pitfalls and solutions. *Bone Marrow Transplantation*, *57*(9), 1428-1434. doi:10.1038/s41409-022-01740-4

Version:	Publisher's Version
License:	Licensed under Article 25fa Copyright Act/Law (Amendment Taverne)
Downloaded from:	https://hdl.handle.net/1887/3564259

Note: To cite this publication please use the final published version (if applicable).

EBMT

ARTICLE

Analysis of survival outcomes in haematopoietic cell transplant studies: Pitfalls and solutions

Liesbeth C. de Wreede[1,2]✉, Johannes Schetelig[2,3] and Hein Putter[1]

© The Author(s), under exclusive licence to Springer Nature Limited 2022, corrected publication 2022

SERIES EDITOR INTRODUCTION

The final article in our Statistics Series by de Wreede and colleagues deals with the important issue of survival analyses in general and in recipients of haematopoietic cell transplants specifically. At first glance analyzing survival should be simple. The endpoint is clear with rare exception, the subject is either alive or dead. Compare this to other less well defined transplant-related outcomes such as who has acute graft-*versus*-host disease (G*v*HD) and of what grade or what is the cause of interstitial pneumonia. There is also the complexity of composite endpoints when one analyzes outcomes such as event-free (EFS) or relapse-free survival (RFS). Here you're either alive or dead. Period. Alas, as it turns out things are not so simple. As the authours point out: *it takes time to observe time*. It is almost never possible to wait long enough for everyone in a study to die. (Some people who are cured by a transplant will outlive their physician and statistician.) Other subjects may not be followed until the end of the study, lost to follow-up or withdraw consent to participate. Often these are non-random events, muddy the water and make what seems a simple analysis of survival not so. Fortunately, de Wreede and colleagues discuss the issues of *informative and non-informative censoring* and time-dependent co-variates. And there are other nasty complexities such non-proportional hazards of death say when initially there is a survival disadvantage to transplants from transplant-related mortality followed in 1–2 years by a survival benefit. They emphasize the danger of considering only Hazard Ratio in this setting. Lastly, the authours discuss how to compare interventions such as conventional therapy *versus* a haematopoietic cell transplant when the endpoint of interest is survival. We think this article will be of considerable interest to readers of BONE MARROW TRANSPLANTATION and suggest you study it carefully. Survival analyses, seemingly simple, are a potential minefield. You don't want to step on one. This article and the entire Statistics Series are available online at https://www.nature.com/collections/ejhigdbeeh.

Robert Peter Gale MD, PhD & Mei-Jie Zhang PhD.

ABSTRACT

The most important outcome of many studies of haematopoietic cell transplants is survival. The statistical field that deals with such outcomes is survival analysis. Methods developed in this field are also applicable to other outcomes where the occurrence and timing are important. Analysis of such *time-to-event outcomes* has special challenges because *it takes time to observe time*. The most important condition for unbiased estimation of a survival curve—non-informative censoring—is discussed along with methods to account for competing risks, a situation where multiple, mutually-exclusive endpoints are of interest. Techniques to compare survival outcomes between groups are reviewed, including the instance where it is unknown at baseline to which group a subject will belong later during follow-up (time-dependent covariates).

Bone Marrow Transplantation (2022) 57:1428–1434; https://doi.org/10.1038/s41409-022-01740-4

INTRODUCTION

An essential outcomes measure in haematopoietic cell transplant studies is survival duration. The eventual efficacy of a treatment, or relevance of a risk factor, is determined by its effect on survival. With certainty, all subjects will eventually die, but beforehand it is unknown when this will take place. The domain of statistics which analyzes these outcomes is survival analysis. Methods from this domain are also suitable for analyzing other outcomes for which it is not only relevant if they occur but also when, such as engraftment, relapse, complications and subsequent interventions. There are many textbooks about survival analysis and clinician-oriented overviews including the *EBMT Statistical Guidelines* that explain many principles also relevant outside the context of transplant studies [1–7]. This topic has also been discussed in the statistical series to which this typescript belongs [8–14]. We aim to give an intuitive explanation of the main concepts of survival analysis relevant for understanding clinical studies in the transplant field while warning about some pitfalls commonly encountered in publications and reports. Typical pitfalls are creating bias by inducing informative censoring, ignoring competing risks, not investigating proportionality of hazards when applying a Cox model and treating time-dependent information as if it were known at the start.

[1]Department of Biomedical Data Sciences, LUMC, Leiden, the Netherlands. [2]DKMS, German Bone Marrow Donor Center, Dresden, Germany. [3]Medical Department I, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany. ✉email: l.c.de_wreede@lumc.nl

Received: 26 April 2022 Revised: 1 June 2022 Accepted: 8 June 2022
Published online: 27 June 2022

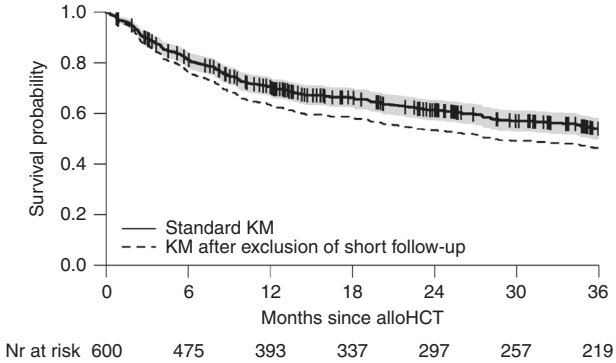

Fig. 1 Kaplan-Meier curves for overall survival. Standard Kaplan-Meier curve (solid line, vertical bars indicate censored observations) and the Kaplan-Meier estimate based on exclusion of patients with a follow-up of less than 36 months and still alive at the end of their follow-up (dashed line). The grey zone indicates the 95%-pointwise confidence intervals. Below the figure numbers at risk at different time points are given, indicating how many patients are still in the risk set (not died and not yet censored, correct analysis).

Time-to-event endpoints are always expressed by time-dependent probabilities, *e.g.*, 1-year survival probability after start of treatment was 85% or 2-year probability of being alive without a new relapse in a patient who had achieved a remission (relapse-free survival, RFS) was 70%. Statements about percentages are uninterpretable without time information: at the starting time the probability of being event-free, such as being alive or being alive without relapse in the above examples, is 100%. Since the life span of all living beings is finite, the probability to be alive of any cohort will be zero if we wait long enough.

As a measure of uncertainty in the estimates the 95% Confidence Interval (CI) is usually added. The CI has to be interpreted by considering the dataset at hand as one possible sample from the population of interest. If repeatedly new, random samples were drawn from this population, the true value of interest *(e.g.*, 1-year survival probability) would be contained in 95% of the CIs. This definition implies in 5 percent of studies, the true value will lie outside the 95% CI. This reflects the common practice in testing for significance where a significance level of 5 percent is used implying a 5 percent probability of false rejection of the null hypothesis.

The data file

We discuss pitfalls and solutions by analyses of a real data file from the EBMT but slightly changed and abstracted to better explain the general principles. It contains information about 600 subjects with a certain kind of leukaemia who received standard or experimental treatment (*e.g.*, different kinds of conditioning or graft-*versus*-host disease (GvHD) prophylaxis). Subjects were followed since transplant. Two kinds of failure are considered: death and relapse of the disease (all patients were in remission at transplant). Analyses were performed in the freely available software R, version 4.1.2 (https://cran.r-project.org/), with packages 'survival' [15] and 'prodlim' [16]. Code is available from the corresponding author on request.

We will also give some examples from published registry-based EBMT studies.

Censoring

The main difficulty in estimating survival probabilities is that it "*takes time to observe time*". It is almost never possible to wait long enough until all patients in a study have experienced the relevant event. Moreover, there are often patients who have not been followed until the end of the study, *e.g.*, due to moving or withdrawing consent. These patients are *lost to follow-up*. For all subjects with a follow-up greater than zero we know that during a certain time interval they could have experienced the event of interest and that this did not occur during the time they were followed. However, we do not know what happened to them afterwards. This phenomenon is called *censoring*. Typically, this interval is different for all patients in the data file because each patient starts the treatment and finishes follow-up at a different moment.

Censored observations cannot be ignored in analyses because that would create a *bias*: a systematic difference between the real outcome and the outcome estimated in the analyses. This bias is caused by inclusion of patients who died relatively early with simultaneous exclusion of those followed during approximately the same time interval but still alive thereafter. This leads to an over-estimation of the probability of dying (Fig. 1). The black curve shows the correct estimate of the survival duration calculated by means of the Kaplan-Meier method in which all censored observations are indicated with a vertical bar [3, 17]. The dashed curve displays the incorrectly estimated probabilities where all patients censored within three years after transplant are removed from the data file. The difference between the two curves indicates the (estimated) bias. Confidence intervals are added to the correct curve to indicate the precision of the estimated probabilities. Moreover, numbers of patients at risk over time are added. These numbers are reduced by events and by censoring. No reliable curves can be shown beyond the time when only 7–10 patients are at risk.

The proper estimation of outcomes in survival analysis is thus made complex because all data of all patients must be integrated, *i.e.*, the time during which they were followed and whether or not they experienced the event of interest. This is managed appropriately by means of the three most prominent methods in this field: [1] the Kaplan–Meier method for estimation of probabilities; [2] the log-rank test for the comparison of survival curves; and [3] Cox's proportional hazards model to estimate the impact of several risk factors at once on outcomes [18]. One assumption essential for these methods is: patients still in follow-up at later timepoints share the same risk profile with those who are censored at the moment of their end of follow-up. In other words, if the follow-up of a patient ends, *e.g.*, at one year after transplant, then the subsequent survival probabilities for this patient are equal to those of someone still in follow-up. This assumption makes it possible to let the patients with longer follow-up represent the censored patients. It is a reasonable assumption if the reason for censoring is independent of the prognosis of the patient, for instance in case of end of study. However, if for instance patients with a good prognosis have briefer follow-up because they stop visiting their clinicians for check-ups, this assumption does not hold, and the Kaplan-Meier method does not give valid estimates. Other methods are not capable of solving this problem either because no data are available after their last follow-up for the censored patients. It is therefore very important to restrict loss to follow-up as much as possible in a study and in particular to counteract mechanisms that make censoring more likely for some categories of patients compared with others.

Unfortunately, registries to which centers contribute on a semi-voluntary basis such as the EBMT registry suffer from loss to follow-up. For example, after autotransplants many patients are no longer monitored at the transplant center resulting in incomplete outcomes data. If the patients with longer follow-up are not representative of the entire transplant population this results in bias in outcomes analyses. EBMT is committed to remedy this problem in the context of a benchmarking process [19, 20].

A particular issue in long-term outcomes studies is that by definition only patients transplanted long ago can contribute information about the long term. Still, the data of more recent patients help better estimate the first part of the trajectory. They are censored before the end of the study. In a strict sense, this can be considered non-informative censoring only if the selection of transplant patients in more recent years resembles earlier

selection and if outcomes are comparable over the years. In an EBMT study about long-term outcomes after allogeneic hematopoietic stem cell transplantation (alloHCT) for patients with myelodysplastic syndromes and secondary acute myeloid leukaemia, patients transplanted before 2013 were included in a dataset closed in December 2016. Median follow-up of survivors was 4.4 years. The authors investigated completeness of follow-up and its impact on outcomes and concluded that 10-year outcomes could be reported reliably [21].

Informative censoring. In practice, it is not always feasible to prevent loss to follow-up, especially in studies based on retrospective data collection. Sometimes, however, investigators decide to censor data which can have tremendous detrimental consequences for their analysis. For example, they might want to compare efficacy of a standard and experimental intervention as initial therapy. After a posttransplant leukaemia relapse many people receive an intervention such as a donor lymphocyte infusion which can be associated with excess mortality. The investigators are not interested in second-line treatment and therefore they reason that they have to censor patients when they relapse posttransplant or as soon as they are treated for relapse. This approach violates the aforementioned assumption of non-informative censoring. Outcomes of the patients with a worse prognosis because they relapse are systematically removed from the analysis. The resulting survival curves have no real meaning, they represent the survival probabilities in a world where people with leukaemia cannot experience a relapse, an unrealistic perspective. Nor can they receive a salvage intervention. Figure 2 illustrates the impact of censoring of patients at relapse and shows the upward bias in the wrongly constructed curves. In this illustration the difference between the two correct and two wrong curves is similar but if the intermediate event (an event taking place between start of study and death) is much more frequent in one group than the other, the bias in the two curves will have a different size and the comparison between the groups can be far off. This wrong method is still often used in practice resulting in biased curves and hampering comparison between studies [22]. A typical example is the analysis of outcomes of leukaemia patients in studies where some of the patients received alloHCT as consolidation therapy instead of post-remission chemotherapy. If the researchers are mainly interested in the non-cellular part of the therapy and censor the patients at alloHCT, they create bias in the survival analysis starting at first treatment. The direction of the bias depends on whether transplant recipients are a good- or poor-risk subset.

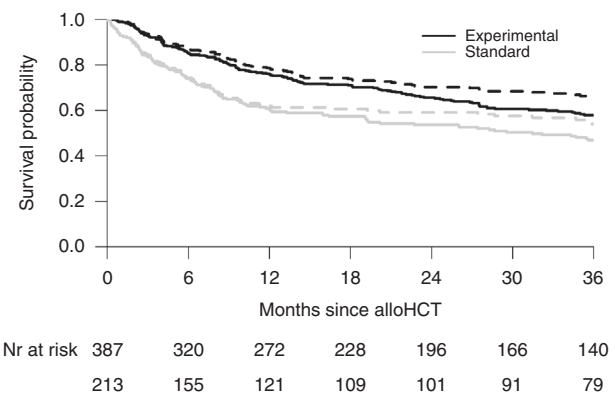

Fig. 2 Correct and incorrect Kaplan-Meier curves. The standard (correct) Kaplan-Meier curves (for experimental and standard treatment) are indicated by the solid lines; the biased Kaplan-Meier curves where patients with a relapse have been censored at the moment of relapse are indicated by dashed lines. Below the figure, correct numbers at risk for the two groups are given.

It follows that censoring should not be induced by an intermediate event such as second-line treatment because this almost always results in informative censoring. When evaluating the impact of an intervention on survival, all deaths have to be included in the analysis because they all are all a direct or indirect consequence of the initial intervention. When the outcome of interest is only posttransplant deaths, but without second-line treatment having taken place, a different method needs to be used (see below).

Methods for competing risks. The methods described above can be extended in such a way that we do not consider a single outcome – death or relapse-or-death (whichever comes first) in the case of relapse-free survival—but multiple outcomes, for instance relapse and death before relapse (non-relapse mortality, NRM) as two separate outcomes. This can be done by means of methods for competing risks: they compete with each other because the occurrence of one of them makes it per definition impossible that the other takes place [7, 23, 24]. In order to see the impact of two treatments on disease activity, for instance, the cumulative incidence of relapse can be calculated, that is, the time-dependent probability of relapse. Patients who die before disease recurrence, for instance due to GvHD or infections, are not being censored, but included in the calculation of NRM. This name is preferable to treatment-related mortality, since causes of death of patients who die before disease recurrence can also be non-HCT related, such as comorbidity. In this way, the two components of the end of RFS —relapse, and NRM—are separated.

These competing risks methods, which take away the biases of naïve methods as the Kaplan–Meier or crude proportions, must be used for all non-fatal outcomes after transplantation, such as engraftment or GvHD, since in the context of alloHCT, death before the event of interest is always a non-negligible competing risk. In a recent large EBMT study about acute GvHD, the cumulative incidence of aGvHD grade II-IV until 100 days after alloHCT in different transplantation periods was reported, taking death before occurrence of aGvHD grade II-IV as competing event [25].

An important advantage of this approach is that it also enables to analyze the impact of treatments on each of the components. It is very well conceivable that the experimental treatment has no net effect on RFS but nevertheless works in a different way from the standard treatment: larger efficacy (*ergo*, smaller relapse probability) can concur very well with larger toxicity (*ergo*, higher NRM probability). Figure 3 shows a situation where the benefit of the new treatment is mainly caused by decreased NRM.

Comparison of treatments

Figure 2 shows the estimated survival curves for the standard and the experimental treatment. The estimated percentages for the experimental and standard treatment, respectively, are 76.1% and 60.4% at 1 year, and 57.9% and 46.9% at 3 years after transplant. Can we conclude from these numbers that the survival after the experimental treatment is better than after the standard treatment? In the first place, systematic differences must be distinguished from random fluctuations due to the fact that the analyses are based on a sample and not on the whole population. This is managed by means of a statistical test. In the second place, the survival curves consist of estimated survival probabilities for both groups at many different points in time. This could potentially lead to a multitude of tests whereas in principle the interest is in the comparison of the survival curves as a whole. The test most commonly used for comparing survival curves is called the *log-rank test* since it takes into account the order of the events in the groups that are compared. In Fig. 2, the p-value of the log-rank test is 0.0002, indicating a highly significant difference between the two curves.

With small sample sizes (less than 10 individuals in one group), the log-rank test is not reliable anymore and can erroneously

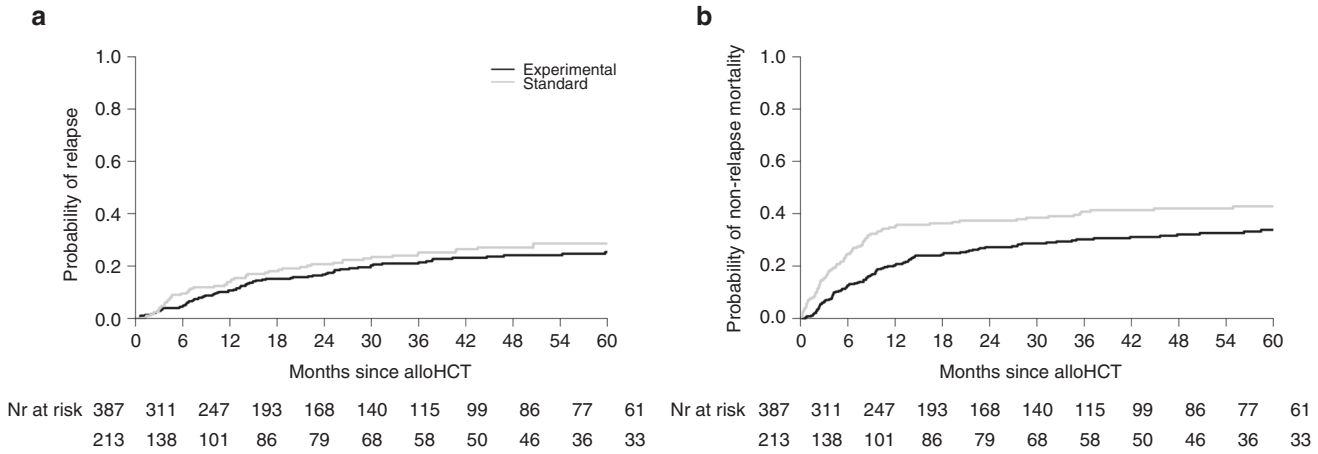

Fig. 3 Cumulative incidence curves for relapse (a) and non-relapse mortality (b) for experimental and standard treatment. The experimental treatment seems to be mainly effective in reducing the NRM probability.

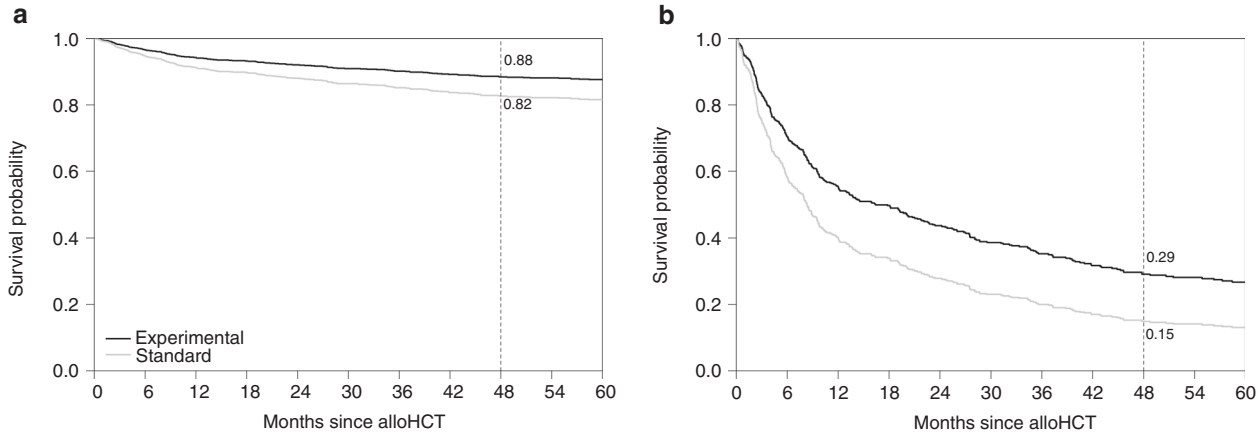

Fig. 4 Survival curves for two treatments with a Hazard Ratio of 1.55. This HR is equal to the HR in Fig. 2, solid lines. Compared to Fig. 2, the baseline hazard has been divided by 5 (**a**) and multiplied by 2.5 (**b**). Point estimates at 48 months are given, indicating that the same HR can lead to different predicted (model-based) probabilities and different differences between these for the two treatments.

indicate a significant difference between groups. A better alternative in this setting are exact permutation tests [26–28].

Proportional hazards models. The log-rank test addresses the question whether survival is the same for two treatments. If it is shown to be different, the question becomes: how large is the effect of the treatment? One way to answer this question is to go back to the estimated survival probabilities at 3 years. The difference between the 46.9% of the standard and the 57.9% of the experimental treatment equals 11.0% and can be seen as an estimate of the effect size of the treatment on survival at 3 years. The advantage of quantifying the effect of the treatment in this way is that it is simple to calculate the so-called number needed to treat (NNT), the number of patients one would need to treat with the experimental treatment (compared to standard) to prevent one death within 3 years. This NNT equals 9 in this case (100 divided by 11). The disadvantage of this approach is that the choice of 3 years is somewhat arbitrary; at 1 year the difference between the survival probabilities would be 15.7% and at 5 years it would be 12.1%. It would be easier to have a single effect measure for the whole follow-up period.

A popular way to quantify the effect of a treatment on survival is through the *hazard ratio*. The *hazard rate*, usually referred to as hazard, describes the probability of death in the next short time interval (for subjects that are alive at that time). The hazard rate varies over time; generally, the hazard rate will increase in the long term because of ageing: the average eighty-year-old has a higher probability to die than the average forty-year-old. But there are other biological mechanisms that can cause the hazard rate to vary over time. For instance, an aggressive treatment like alloHCT will cause the hazard rate to be increased in the initial period after transplantation. Mathematical formulas exist to express the survival probabilities in terms of the hazards and vice versa; in other words, the hazard contains the same information as the survival curve, although its perspective is different.

When we compare two treatments, the hazard rates of both treatments will vary over time. The *proportional hazards (PH) model* makes an important assumption, namely that the *hazard ratio (HR)*, *i.e.*, the ratio between the two hazard rates, remains constant. It is a model, which means that a certain structure is imposed on the data. In the above example (see Fig. 2) the HR of the standard treatment compared to the experimental treatment equals 1.55, with a 95%-confidence interval running from 1.23 to 1.94. This means that at each moment in time the instantaneous risk of dying for the patients receiving the standard treatment is 1.55 as high as that of the patients receiving the experimental treatment. The HR can also be interpreted as an average risk ratio over time. When a HR is reported it has to be clear what is compared to what. The hazard ratio of the experimental treatment with respect to (or *versus*) standard is 0.65 (=1/1.55).

The hazard ratio is a multiplication factor of the hazard, not of the survival or death probabilities. A HR of 1.55 does *not* mean

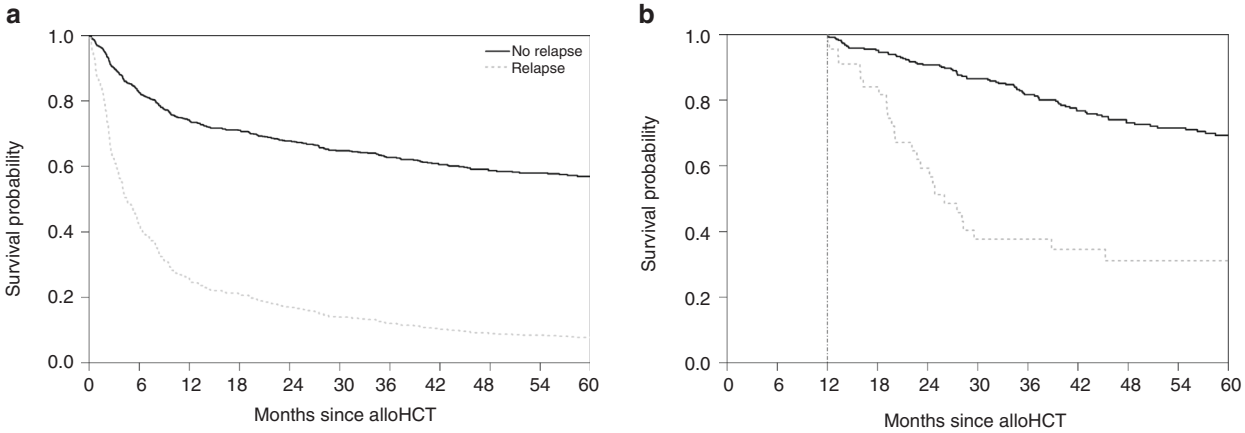

Fig. 5 Analyses for the impact of relapse on death. a Cox model-based survival curves showing predicted survival probabilities for patients with a relapse at time of HCT *vs*. patients who never relapse; **b** survival curves based on a landmark analysis at 12 months after alloHCT showing survival probabilities for patients alive at the 1 year landmark who have experienced a relapse before that time *vs*. those who have not.

that the probability to die within 3 years in the group receiving the standard treatment is 1.55 times as high as in the group receiving the experimental treatment. A HR of 1.55 can manifest itself in various ways in the two corresponding survival curves. The survival curve of the experimental treatment has to lie above that of the standard treatment, but the *baseline hazard*, the hazard of the reference group, determines how far apart the curves are. With a high baseline hazard the survival of the reference group is low and the survival of the other group is even lower; with a low baseline hazard, the survival of the reference group is high (close to 1), and the survival of the other group will be worse, but possibly still close to 1. Figure 4 shows two comparisons of two survival curves, both with a HR of 1.55, as in Fig. 2. The baseline hazard of Fig. 2 has been decreased (Fig. 4(a)) and increased (Fig. 4 (b)). The curves look quite different, but in all three cases the HR is 1.55.

Non-proportional hazards. The crucial assumption of the proportional hazards model is that the ratio of two hazard curves (both varying over time) is constant. This means that we assume that the effect of treatment immediately after alloHCT is the same as later in the follow-up. This assumption turns out to be satisfied reasonably often in practice, especially when the follow-up is not too long. A number of situations exist where the assumption on proportional hazards is not realistic (see Van Houwelingen and Putter, Chapter 5 for a discussion on mechanisms leading to violation of the PH assumption [29]). An important example is the comparison of a standard treatment with a more aggressive treatment (like chemotherapy and alloHCT in an analysis starting from diagnosis), which may be expected to be disadvantageous in the beginning but may result in better outcome in the long term. The HR of experimental with respect to standard will be higher than 1 in the beginning (because of the higher post-transplant mortality) and below 1 later (when relapse and deaths resulting from relapses have been prevented). When a PH model is used in such a situation, the estimated HR (under the incorrect PH assumption) will be an average of the HRs over time [30]. The result of the analysis will depend on the length of the follow-up; with shorter follow-up the experimental treatment will come out worse than with longer follow-up. Hence it is important to check the PH assumption and to specify the timing of the analysis.

With longer follow-up, the assumption becomes more questionable for several reasons: risk factors become less relevant over time, either because they are outdated (*e.g.*, performance status at diagnosis might not be very informative for performance status 5 years later), due to selection (the patients who survived with a poor molecular marker for 5 years are a 'lucky' subset of all patients with this marker at diagnosis) or because other risk factors become more relevant long-term, such as those associated with accelerated ageing and secondary malignancies. These changes in impact can be accommodated within the Cox model by modelling time-dependent HRs or by stratification (estimating baseline hazards separately for different groups while assuming the effect of other covariates to be the same). Also other models exist that are more suitable for long-term outcomes, such as cure models [31] and models incorporating general population mortality (relative survival) [32, 21].

Time-dependent covariates

So far, we have compared two treatments in a PH model. Treatment serves as a so-called *covariate* in the regression model. Often researchers are interested in the effect of a covariate whose value changes over time, a so-called *time-dependent covariate*. Important examples are intermediate events like GvHD and relapse and post-alloHCT treatments like donor lymphocyte infusions. Care is needed in the analysis of the effect of such time-dependent covariates. If the analysis does not properly take into account the fact that the value of the time-dependent covariate is not constant over time, radically incorrect conclusions can be obtained. As an example, consider the effect of the intermediate event relapse on survival. The first thought for estimating the effect of relapse on survival would perhaps be to use a PH model with relapse as an ordinary covariate, known at baseline. In fact, we are then comparing the survival of two groups, one with and the other without relapse. This comparison is unfair, however, because the patients in the relapse group can only have made it into the relapse group provided they have lived long enough to experience this relapse. If they had died before, they would have ended up in the no relapse group. In some sense the patients in the relapse group are immortal until the time they experienced the relapse and the time until relapse contributes to the survival of the relapsed patients although they were still in remission then. The bias caused by this incorrect analysis is known as *immortal time bias* or *guarantee time bias* [33, 34]. If we nevertheless performed this (wrong) analysis, we would obtain a hazard ratio of 1.08 for the relapse group, compared to the no relapse group, with a 95%-confidence interval from 0.84 to 1.38 ($p = 0.58$). The (incorrect) conclusion would be that relapse does not influence survival.

There are several correct methods for performing such analyses. The first option is a so-called *proportional hazards model with time-dependent covariates* [35, 36]. This model uses at each point in time the relapse status at that time (which indicates if the patient has experienced a relapse before that time point) for the

comparison of the hazards. Applying this time-dependent Cox model gives a hazard ratio of 4.53 for relapse with respect to no relapse. That means that if we compare, at each moment in time, two (living) patients, one of which did and the other did not (yet) have a relapse, then the patient with relapse has a 4.53 times higher instantaneous risk of dying, compared to the patient without relapse. Figure 5(a) shows the model-based curves, comparing the outcomes of two reference patients: one who never experiences a relapse and one who has a relapse immediately after alloHCT.

The second option is *landmarking* [34, 37, 38]. At a prespecified moment in time, the landmark time point, the patients alive and under follow-up at that time are considered, and a comparison is made between the group that had experienced a relapse before the landmark and the group that had not. Note that some of the patients with no relapse before the landmark might experience a relapse afterwards; such relapses occurring after the landmark are considered irrelevant for the landmark analysis. With a landmark time point set at one year, this comparison results in a hazard ratio of 3.93 for relapse with a 95%-confidence interval from 2.56 to 6.03. Figure 5(b) shows the Kaplan-Meier survival curves of the two groups. Both Cox models with time-dependent covariates and landmarking are correct methods but approach the comparison from a different perspective. The time-dependent Cox model gives a summary of the effect of relapse on death over the whole follow-up, while the landmark model limits the comparison to studying the effect of relapses having occurred before the landmark on deaths occurring after the landmark.

An alternative is to start the analysis at the moment of the intermediate event. This is a correct and simple approach, however not suitable for comparing outcomes with and without the event. This approach was followed in the EBMT study about aGvHD where survival outcomes after onset of aGvHD were reported, taking time of aGvHD as new time 0 instead of time of alloHCT [25].

An important example of a question related to a time-dependent covariate is whether, and if yes, when during the course of the disease patients with hematological malignancies should receive a transplant. This question cannot be answered by EBMT data alone since per definition only patients who have received a transplantation enter the registry; data about the selection process between diagnosis and transplantation are lacking (again, patients are "immortal" during the interval diagnosis to HCT). Additional data from disease-based registries are sometimes used to address this question [39, 40]. Especially in a treatment-based registry as is EBMT, opposed to population-based disease registries, selection for the treatment – by clinical characteristics or policy – always plays a role and might prevent generalization to a larger target group.

CONCLUSION

When analyzing time-to-event data and when interpreting results in papers, it is important to be aware of the assumptions underlying the most common techniques: non-informative censoring and proportional hazards. If these assumptions are not fulfilled, the results of the analysis can be very misleading. Another important point is how and when to analyze composite endpoints such as RFS. Composite endpoints are easier to analyze than separate endpoints as analyzed in a competing risks model. However, to understand the mechanisms behind failure, it is very useful to distinguish the different causes of failure and the impact of risk factors on them.

When proportional hazards models are used, we recommend to study not only the hazard ratio, but also the baseline hazard, because only a combination of both gives a good measure of the effect of a treatment or risk factor on outcome.

For further study of the issues discussed here, the referenced books and papers can be used. Especially recommended is the recent overview paper of the STRATOS (STRengthening Analytical Thinking for Observational Studies) topic group Survival Analysis, which explains in more depth the issues discussed in the current paper, including the main mathematical aspects and elaborate example code in R [41]. In case of non-standard studies, consultation of a statistician with experience in survival analysis is strongly recommended.

DATA AVAILABILITY

The data file used in this article can be obtained from the corresponding author on reasonable request.

REFERENCES

1. Iacobelli S, EBMT Statistical Committee. Suggestions on the use of statistical methodologies in studies of the European Group for Blood and Marrow Transplantation. Bone Marrow Transpl. 2013;48:S1–37. Suppl 1
2. Klein JP, Moeschberger ML. Techniques for censored and truncated data. New York: Springer-Verlag; 2003.
3. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89:232–8.
4. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis—an introduction to concepts and methods. Br J Cancer. 2003;89:431–6.
5. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part III: Multivariate data analysis—choosing a model and assessing its adequacy and fit. Br J Cancer. 2003;89:605–11.
6. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part IV: Further concepts and methods in survival analysis. Br J Cancer. 2003;89:781–6.
7. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. Br J Cancer. 2004;91:1229–35.
8. Gale RP, Zhang MJ. Statistical analyses of clinical trials in haematopoietic cell transplantation or why there is a strong correlation between people drowning after falling out of a fishing boat and marriage rate in Kentucky. Bone Marrow Transpl. 2020;55:1–3.
9. Zheng C, Dai R, Gale RP, Zhang MJ. Causal inference in randomized clinical trials. Bone Marrow Transpl. 2020;55:4–8.
10. Hu ZH, Peter Gale R, Zhang MJ. Direct adjusted survival and cumulative incidence curves for observational studies. Bone Marrow Transpl. 2020;55:538–43.
11. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transpl. 2020;55:675–80.
12. Othus M, Gale RP, Hourigan CS, Walter RB. Statistics and measurable residual disease (MRD) testing: uses and abuses in hematopoietic cell transplantation. Bone Marrow Transpl. 2020;55:843–50.
13. Moodie EEM, Krakow EF. Precision medicine: Statistical methods for estimating adaptive treatment strategies. Bone Marrow Transpl. 2020;55:1890–6.
14. Hu ZH, Wang HL, Gale RP, Zhang MJA. SAS macro for estimating direct adjusted survival functions for time-to-event data with or without left truncation. Bone Marrow Transpl. 2022;57:6–10.
15. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox Model. New York: Springer Science & Business Media; 2000. 372 p.
16. Gerds TA. prodlim: Product-Limit estimation for censored event history analysis. 2019. Available from: https://CRAN.R-project.org/package=prodlim
17. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–81.
18. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34:187–220.
19. Snowden JA, Saccardi R, Orchard K, Ljungman P, Duarte RF, Labopin M, et al. Benchmarking of survival outcomes following haematopoietic stem cell transplantation: A review of existing processes and the introduction of an international system from the European Society for Blood and Marrow Transplantation (EBMT) and the Joint Accreditation Committee of ISCT and EBMT (JACIE). Bone Marrow Transpl. 2020;55:681–94.
20. Putter H, Eikema DJ, de Wreede LC, McGrath E, Sánchez-Ortega I, Saccardi R, et al. Benchmarking survival outcomes: A funnel plot for survival data. Stat Methods Med Res. 2022 Mar;09622802221084130.
21. Schetelig J, de Wreede LC, van Gelder M, Koster L, Finke J, Niederwieser D, et al. Late treatment-related mortality versus competing causes of death after

allogeneic transplantation for myelodysplastic syndromes and secondary acute myeloid leukemia. Leukemia 2019;33:686–95.
22. Kim HT, Armand P. Clinical endpoints in allogeneic hematopoietic stem cell transplantation studies: the cost of freedom. Biol Blood Marrow Transpl. 2013;19:860–6.
23. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26:2389–430.
24. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. Int J Epidemiol. 2012;41:861–70.
25. Greinix HT, Eikema DJ, Koster L, Penack O, Yakoub-Agha I, Montoto S, et al. Improved outcome of patients with graft-versus-host disease after allogeneic hematopoietic cell transplantation for hematologic malignancies over time: an EBMT mega-file study. Haematologica 2022;107:1054–63.
26. Latta RB. A Monte Carlo study of some two-sample rank tests with censored data. J Am Stat Assoc. 1981;76:713–9.
27. Kellerer AM, Chmelevsky D. Small-sample properties of censored-data rank tests. Biometrics 1983;39:675–82.
28. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a class of permutation tests: The coin Package. J Stat Softw. 2008;28:1–23.
29. van Houwelingen H, Putter H. Dynamic prediction in clinical survival analysis. Boca Raton: CRC Press; 2012. 250 p.
30. Schemper M. Cox analysis of survival data with non-proportional hazard functions. J R Stat Soc Ser Stat 1992;41:455–65.
31. Felizzi F, Paracha N, Pöhlmann J, Ray J. Mixture cure models in oncology: a tutorial and practical guidance. PharmacoEconomics - Open. 2021;5:143–55.
32. Pohar Perme M, Pavlic K. Nonparametric relative survival analysis with the R Package relsurv. J Stat Softw. 2018;87:1–27.
33. Hsieh PY, Liu CJ, Teng CJ. Immortal time bias in retrospective analysis: comment on "Efficacy and safety of long-term treatment with lenalidomide and dexamethasone in patients with relapsed/refractory multiple myeloma.". Blood Cancer J 2015;5:e283.
34. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1:710–9.
35. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. Annu Rev Public Health. 1999;20:145–57.
36. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med. 2018;6:121.
37. Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes. 2011;4:363–71.
38. Morgan CJ. Landmark analysis: A primer. J Nucl Cardiol. 2019;26:391–3.
39. Brand R, Putter H, van Biezen A, Niederwieser D, Martino R, Mufti G, et al. Comparison of allogeneic stem cell transplantation and non-transplant approaches in elderly patients with advanced myelodysplastic syndrome: optimal statistical approaches and a critical appraisal of clinical results using non-randomized data. PLOS ONE. 2013;8:e74368.
40. Robin M, de Wreede LC, Padron E, Bakunina K, Fenaux P, Koster L, et al. Role of allogeneic transplantation in chronic myelomonocytic leukemia: an international collaborative analysis. Blood (accepted 2022). https://doi.org/10.1182/blood.2021015173.
41. Kragh Andersen P, Pohar Perme M, van Houwelingen HC, Cook RJ, Joly P, Martinussen T, et al. Analysis of time-to-event for observational studies: Guidance to the use of intensity models. Stat Med. 2021;40:185–211.
42. de Wreede LC, Putter H. Valkuilen en oplossingen bij de overlevingsduuranalyse in hematologische studies/Pitfalls and solutions in survival analysis for hematological studies. NTVH. 2017;14:118–24.

ACKNOWLEDGEMENTS

We thank Robert P. Gale (Imperial College, London, UK) and Mei-Jie Zhang (Medical College of Wisconsin, Milwaukee, WI, USA) for their invitation to contribute to their series of papers about statistical topics and we thank Dr Gale for his suggestions about content and wording that have improved the manuscript. We thank Michel van Gelder (MUMC, Maastricht, the Netherlands) and Katharina Schmidt-Brücken (Technical University Dresden, Germany) for their critical review of previous versions of this manuscript. This article is a reworked version of an article in Dutch [42].

AUTHOR CONTRIBUTIONS

LdW and HP designed the study, wrote the manuscript, and analyzed the data. JS critically reviewed the manuscript. All authors read and approved the final manuscript.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Liesbeth C. de Wreede.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.