Bone Marrow Transplantation (2001) 28, 909–915

www.nature.com/bmt

Mini-review

Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part I: Unadjusted analysis

JP Klein[1,2], JD Rizzo[2], M-J Zhang[1,2] and N Keiding[3]

[1]*Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI, USA;* [2]*IBMTR/ABMTR, Milwaukee, WI, USA; and* [3]*Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark*

Summary:

In this paper, we describe modern statistical methods for presentation of the results of studies of bone marrow transplantation. We focus here on 'univariate' or unadjusted techniques to describe the outcomes of such studies. In another paper we will discuss multivariate methods. We discuss the type of data one may have available to make inference about outcomes. We explain the differences between the Kaplan–Meier estimator of the survival function and the cumulative incidence curve, how these curves should be interpreted and when each is the appropriate summary statistic. We discuss the weighted log rank statistic and show how different weights can be used to put emphasis on detecting differences between groups in different time periods. We also present a simple estimate of current leukemia-free survival which is useful in summarizing post-transplant events. *Bone Marrow Transplantation* (2001) **28,** 909–915.
Keywords: competing risks; Kaplan–Meier estimates, weighted log rank tests; current leukemia-free survival

Types of data

Complications of bone marrow transplantation (BMT) and the diseases for which it is performed occur frequently. Therefore death, relapse or progression, graft failure, graft-versus-host disease (GVHD) and infection are among the typical outcomes described in BMT studies. Each of these outcomes happens at varying times during the patient's sojourn after transplant. Most investigators are interested in summarizing these outcomes or in comparing occurrence of these outcomes between groups of patients. An important characteristic of BMT studies is that these outcomes occur at random times. Proper handling of time is critical to performing a correct analysis. Further complicating the analysis is the fact that, for many patients, the event of interest has not occurred prior to the time of last observation. Such patients are said to be right censored.

The first step in any analysis of BMT data is to define a proper time scale and time origin for the data. A misspecification of the time origin can lead to biased estimates of all the outcome probabilities of interest. The time origin is most commonly taken to be the time of transplant, but it may be the time of diagnosis when comparing transplant patients to chemotherapy patients or the time of development of acute GVHD of a particular grade if we want to compare the length of survival after development of aGVHD.

All of the statistics discussed in the following sections are based on estimating, at each point in time, the probability of an event occurring among those who could possibly experience the event at this time. The denominator of this estimated probability must include only those patients who are at risk and not patients in the study who have no potential for experiencing the event at this time. For example, if we wished to study post-GVHD survival and we picked as our time origin the time of transplant, patients who are at risk at a given point in time are the number who are alive after GVHD, not the total number of patients who are alive. Here patients do not become at risk and are not counted in the denominators of our statistics unless they have developed GVHD. In fact, if we wished to compare the survival of patients with or without GVHD, all patients would initially be counted in the non-GVHD arm. Once they developed GVHD they would move to the GVHD arm. Since patients do not move into the GVHD arm until random times after transplant, the data in the GVHD arm is called left truncated. This simple adjustment makes the appropriate correction for the fact that patients in the GVHD arm must live until they get GVHD before they can die in this arm.

A common mistake in transplant studies is made in the comparison of the outcomes of post-transplant therapy, or in comparing outcomes between patients with or without a post-transplant condition. For example, one may want to compare survival between patients with and without GVHD or patients receiving second transplants to those who do not. For the purpose of performing a statistical test or to obtain an estimate of the survival curve, the arm a patient will eventually be on is often assumed to be known at the time of transplant. This incorrect assumption leads to biased and meaningless conclusions. When making comparisons between patients at any given point in time, it is essential that any analysis only uses the information known at that

Correspondence: Dr JP Klein, Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI, USA

same time point, and not any information which becomes available in the future. If we wish to compare survival with and without GVHD, we need to either use a method that moves subjects from the no-GVHD arm to the GVHD arm when they develop GVHD, or we need to treat GVHD as a time-dependent covariate in a Cox model.[1,2] Note that if we want to compare planned tandem transplants to single transplants, then the groups should be assigned on an intention-to-treat basis. Here the tandem arm is those patients who, at the time of first transplant were anticipated to receive a planned second transplant regardless of whether they eventually received the transplant.

Transplant outcomes typically fall into one of two types which require different analysis techniques. The first type of data is survival data. The outcome is a simple event which occurs at a random time. Examples are death and treatment failure (death or relapse). Each patient either experiences the event or is censored at the last time we have them under observation without the event having occurred. This type of data is presented as a Kaplan–Meier survival curve.

The second type of data is competing risk data. Competing risks must be considered when the occurrence of one event precludes the occurrence of another event. The classical examples are relapse and death in remission. Here, as discussed in the next section, the proper summary curve is the cumulative incidence curve. For competing risk data, a cumulative incidence curve should be produced for each of the competing risks. Table 1 shows some of the common competing risks which are studied in transplant trials. Note that the last entry, death from primary disease, requires an assessment of the cause of death for each patient.

In the next sections, we discuss in more detail the difference between the Kaplan–Meier and cumulative incidence curves, focusing on the interpretation of these curves. We then examine how one can perform unadjusted comparisons between outcomes in transplant studies using a weighted log rank test. This test can put emphasis on different intervals in time where we are most interested in detecting differences in outcome between groups. Finally, we present a technique for evaluating post-transplant therapies, the current leukemia-free survival curve.

Kaplan–Meier *vs* cumulative incidence curves

As indicated in section 1, the easiest outcomes to analyze are simple events that occur at a random time. The basic example is death, but sometimes a simple event 'treatment failure' is defined as 'death or relapse, whatever comes first'. The Kaplan–Meier[3] survival curve is at a time t since the time origin is a nonparametric estimator of the probability of survival past t (for treatment failure as endpoint, to be interpreted as relapse-free survival). For this estimator, censored observations must come from a mechanism unrelated to the simple event or the estimate is biased.[4] Typically censoring is due to a short study time or to the patient being lost to follow-up. Details of the construction of this estimate, estimates of the standard error[5] and the construction of pointwise confidence intervals and confidence bands[6] can be found in Klein and Moeschberger[2] or Klein.[7] Two Kaplan–Meier curves may be compared using nonparametric tests, of which the log rank test is the most common (see next section).

To illustrate these techniques and those discussed in Part II, consider a subset of data from an IBMTR study of alternative BMT donors for leukemia.[8] The data considered consist of patients transplanted for chronic or acute leukemia with either an HLA-identical sibling donor ($n = 1224$), an HLA matched unrelated donor ($n = 383$) or a 1-HLA antigen mismatched unrelated donor ($n = 108$). Table 2 shows 6-month and 2-year estimates of disease-free survival and 95% confidence intervals based on a log–log transformation.[6]

When several endpoints are considered simultaneously, these tools require more care. We shall take as our basic example two so-called competing risks, of relapse and of death in remission. These events are mutually exclusive: the occurrence of one precludes occurrence of the other. It is straightforward to define the short-term risk that a patient alive and in remission at time t relapses tomorrow (the

Table 2 Estimates of leukemia free survival

	Disease-free survival (95% confidence interval)	
	6 Months post transplant	*2 Years post transplant*
HLA-identical siblings	0.7688 (0.7423–0.7929)	0.5755 (0.5463–0.6035)
HLA-matched unrelated	0.5012 (0.4486–0.5515)	0.3372 (0.2873–0.3877)
1-Antigen mismatched unrelated	0.4167 (0.3223–0.5083)	0.2294 (0.1520–0.3165)

Table 1 Competing risks outcomes

Outcome	*Competing risk*	*Time variable*
Relapse	Death in remission	Time to death or relapse whichever is first
Death in remission	Relapse	Time to death or relapse whichever is first
Progression	Death without progression	Time to death or progression, whichever is first
Acute GVHD	Death without aGVHD, relapse, second transplant	Time to aGVHD, death, relapse, second transplant, whichever is first
Chronic GVHD	Death without cGVHD, relapse, second transplant	Time to cGVHD, death, relapse, second transplant, whichever is first
Engraftment	Death without engraftment	Death, engraftment, whichever is first
Death from disease	Death due to other causes (eg infection, GVHD)	Death time

relapse hazard) or that this patient dies (which will in practice be in remission) tomorrow, the death (in remission) hazard. The sum of these two rates is the treatment failure hazard rate.

The more difficult concepts in competing risk analysis arise when it is desired to define and estimate longer-term risks, such as the probability that a patient will relapse during the next year. A first attempt, unfortunately quite commonly seen in practice, consists of calculating one minus the Kaplan–Meier estimate of relapse, treating deaths in remission as censored observations. This however estimates relapse not in our actual world, but in a counterfactual world where it is impossible to die. This quantity is hard to interpret and provides an overestimate of the chance of relapsing in the real world where patients may die prior to relapsing.

The more appropriate estimate takes proper account of the possibility that the patients are at risk not only for the event of primary interest (here: relapse), but can also be removed from possible relapse because of competing events (here: death in remission). This estimate is the cumulative incidence curve.[9–11] This curve is a function of both the relapse and treatment-related mortality hazard rates and as such is affected by changes in either rate.

Figure 1 shows the (1-Kaplan–Meier) estimate and cumulative incidence curve estimate of relapse for the HLA-identical sibling cohort. It is seen that the cumulative incidence is lower, reflecting the fact that it predicts relapses in this world where some patients die in remission, whereas (1-Kaplan–Meier) attempts to also predict when patients who actually die would have relapsed.

At each time-point, any patient will be in one of three mutually exclusive and exhaustive states: alive and relapse-free, alive and relapsed, or death in remission. The probabilities of these three states have to sum to one (=100%), which as seen in Figure 2, is the case if the cumulative incidence method is used, while this would not hold if (1-Kaplan–Meier) estimates were used.

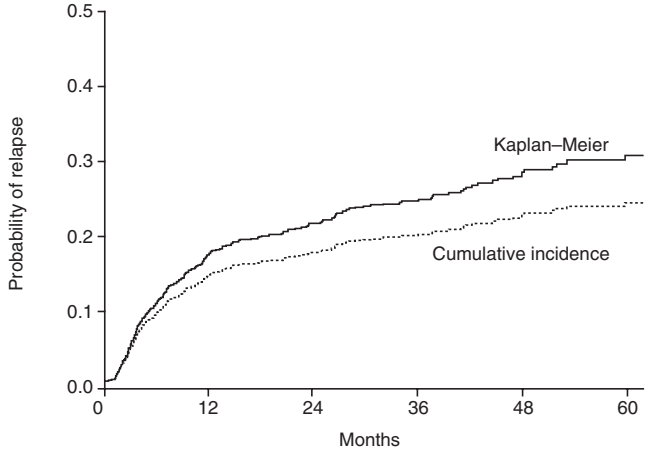

Figure 1 Probability of relapse for HLA-identical siblings.

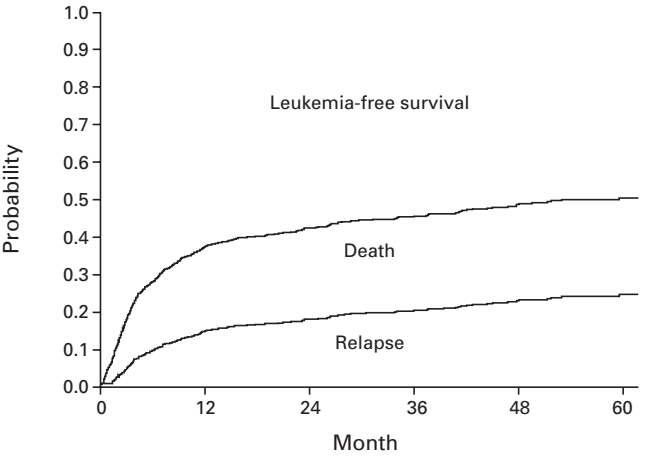

Figure 2 The height of the lower curve is the relapse cumulative incidence at time *t*. The distance between the top curve and the lower curve at *t* months is the relapse-free mortality cumulative incidence at this time. One minus the top curve is the leukemia-free survival probability at *t* months after tranplant. Estimates are for HLA-identical siblings.

Unadjusted comparisons of outcomes between groups

In most studies a comparison of outcomes between two or more groups of patients is of interest. These groups may be determined, for example, by the preparative regimen, the donor cell source, the type of transplant (autologous *vs* allogeneic), or by some other factor. In this comparison it is critical that group membership be known at the time of entry into the study cohort. Studies where the groups are determined by post-transplant events, such as comparing survival in patients with and without chronic GVHD must be handled by other means. These techniques for comparing groups defined dynamically are discussed in our companion paper on regression models for BMT data.

The methods we present here are for unadjusted comparisons between groups where membership is known at onset. We are assuming that there is no other factor which has a differential (confounding) effect on outcome and that other possible prognostic factors are balanced between treatment groups. This should be the case in a randomized trial. In many retrospective studies this may not be a valid assumption. In that case a regression approach, as discussed in our forthcoming paper on regression methods in BMT studies, is more appropriate.

There are three approaches to comparing outcomes in BMT studies. The first is used in making unadjusted comparisons of events, such as engraftment, which occur shortly after transplant. Here one must fix in advance, a time period during which the event will be observed. Patients who experience the event in this window are considered a 'success' and patients who are under study for the entire period but do not experience the event are a 'failure'. Patients who are not followed for the entire time period are not evaluable for this event and are dropped from the study. For example, if we were to compare 100-day mortality rates between patients given an allograft or an autograft we would classify a patient as an event if they die in the first 100 days and as a nonevent if they are alive at 100 days. Patients with less than 100 days of follow-up are not evalu-

able for this event and are dropped from consideration. Comparisons between groups can be made by standard $2 \times k$ chi-square tests, where k is the number of groups. The disadvantage of this approach is that patients not observed for the complete time interval cannot be evaluated for the comparison – a frequent occurrence.

A second type of test is one where we wish to compare survival (or disease-free survival) or cumulative incidence curves (eg relapse or death in remission) at fixed points in time between groups. Here, we compute the estimated probabilities using either the Kaplan–Meier or cumulative incidence curve for each of the k groups, along with the standard errors of these estimates. These estimates are independent so a test analogous to the one way analysis of variance based on a simple quadratic form[12] in the estimates can be constructed. For the comparison of two groups, the test statistic is $Z = (p_1 - p_2)/(V(p_1) + V(p_2))^{1/2}$, where p_1, p_2 are the two point estimates and $V(p_1)$, $V(p_2)$ are the two variance estimates. This Z score should be compared to the standard normal probability table to find the P value. One hundred-day mortality can be compared between groups by this technique using the Kaplan–Meier estimator at 100 days. Since this method uses all the patients and does not discard cases with less than 100 days of follow-up it is more efficient than the chi-squared method.

When one reports, in a tabular form, estimates of survival or competing risk probabilities by groups at fixed times, these are the most appropriate P values to report. However, when probabilities are compared at several time-points then a correction to the significance level needs to be made to account for the additional 'looks' at the data. A 'Bonferroni' correction ensures the overall chance of making an incorrect decision is no more than α if we make each comparison at an (α/K) level, where K is the number of comparisons we are making.

As an example, consider the reported pairwise P values for comparing disease-free survival at 6 months and 2 years for the three donor groups. These P values are given in Table 3. Here, we are making six pairwise comparisons so these P values should be compared to $0.05/6 = 0.0083$ to determine significance to ensure our chance of a false inference is at most 5%. We see that there is strong evidence that the HLA-identical sibling donor transplants do better than either unrelated group, but, even though the 2-year comparison between the two groups of unrelated donor transplants is less than 0.0500, using our corrected cut-off of 0.0083, this difference is not statistically significant.

The third type of test is a (weighted) log rank test. This class of tests compares the entire survival experience for patients in the groups by comparing their hazard rates. The log rank test is computed by comparing the observed number of events for each treatment with the number of events one would expect to see, if there were no difference in the event rates between the groups at each event time. Because this test compares outcome over the whole time interval, it may not adequately detect important differences between groups which occur either early or late in the interval. A weighted log rank test uses a time-dependent weight function to prove greater importance to departures from the null hypothesis of equality of rates at different points in time. A constant weight (over time) leads to the usual log rank test. Weight functions which place more importance on early differences in the groups are Gehan's test[13] or Tarone and Ware's test.[14] These three tests are available in many statistical packages. A more general class of weight functions is due to Fleming and Herrington.[15] This class uses as a weight $S(t)^p\ (1 - S(t))^q$, where $S(t)$ is the pooled Kaplan–Meier estimator at time t (note here we use the Kaplan–Meier estimator for competing risk data since it is a weight function only). When $p = q = 0$, this is the log rank test. When $p = 1$ and $q = 0$, more weight is given to early departures, while when $q = 0$ and $p = 1$ more weight is given to late departures.

Table 3 Pairwise P values comparing disease-free survival probabilities

	6 Months post transplant	*2 Years post transplant*
HLA-identical siblings *vs* 1-antigen mismatched *vs* HLA-matched unrelated	<0.0001	<0.0001
HLA-identical siblings *vs* HLA-matched unrelated	<0.0001	<0.0001
HLA-matched unrelated *vs* 1-antigen mismatched unrelated	0.1220	0.0300

The choice of the weight function must be made before looking at the data using only clinical expectations for the outcome. For survival data, for example, the Wilcoxon test may be best used if we are primarily interested in determining whether early deaths (most likely treatment-related deaths) are different in the two groups. A Fleming and Herrington test with $p = 0$ and $q = 1$ may be the test of choice if we are interested in determining if the plateaus of the treatment curves are different.

Summarizing post-transplant therapy – the current leukemia-free survival probability

In the previous sections we focused on transplant outcomes such as relapse or death where their occurrence causes the

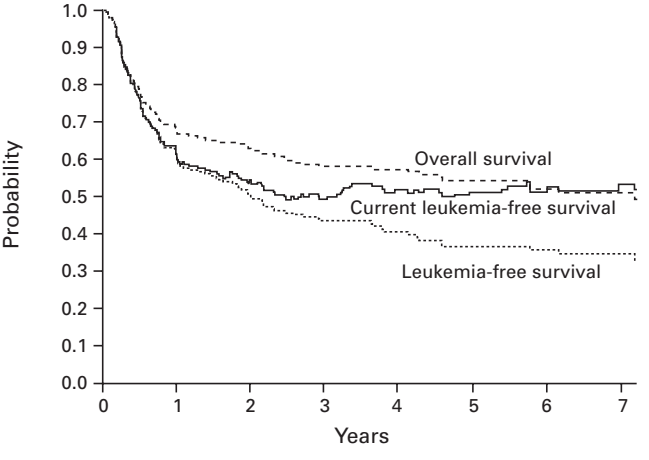

Figure 3 Estimates of current leukemia-free survival, leukemia-free survival and overall survival.

transplant to be considered a failure. We have considered summarizing therapies performed prior to or just after transplant. All the curves we present are proper survival curves that are non-increasing, or in the case of cumulative incidence, non-decreasing.

For disease-free survival the event is 'treatment failure' defined as death or relapse. New therapies, such as donor lymphocyte infusions,[16–20] have been used which can successfully induce a second, durable remission in relapsed patients. To summarize this procedure a new measure of outcome,[21,22] called the current leukemia-free survival (CLFS) has been proposed. Note, as opposed to other suggestions for an estimate of CLFS,[20] this estimator is statistically valid and uses only the information on a patient's status available at a given point in time.

The estimate can be constructed by adding to the usual Kaplan–Meier disease-free survival estimate (event death or first relapse) the difference between a Kaplan–Meier estimator (S_2) where the event is death or second relapse and a Kaplan–Meier estimator (S_3), where the event is death prior to second remission. The difference, $S_2 - S_3$ is the estimated probability of being alive in second remission. This estimate may go up or down reflecting that patients may be in a 'good' health state in different periods and a 'poor' (relapsed) state between visits to the remission state.

To illustrate this consider a study of 189 CML patients reported in Craddock *et al*.[20] Depicted is the overall survival function, the usual disease-free survival function and the current leukemia-free survival function. Note that the current leukemia-free survival function increases, which reflects patients obtaining a second remission. Current leukemia-free survival is applicable to situations where therapy delivered after relapse following a bone marrow transplant can produce a meaningful (prolonged) disease-free survival interval.

Acknowledgements

This research was supported by grant R01-CA54706–07 from the National Cancer Institute.

References

1 Cox, DR. Regression models and life-tables (with discussion). *J Roy Stat Soc* 1972; **B34**: 187–220.
2 Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag: New York, 1997.
3 Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53**: 457–481.
4 Klein JP, Moeschberger ML. Independent or dependent competing risks: does it make a difference? *Comm Stat Simul* 1987; **16**: 507–533.
5 Greenwood M. The natural duration of cancer. In: *Reports on Public Health and Medical Subjects*. His Majesty's Stationery Office: London, 1926, 33: 1–26.
6 Borgan Ø, Liestøl K. A note on confidence intervals and bands for the survival curve based on transformations. *Scand J Stat* 1990; **17**: 35–41.
7 Klein JP. Statistical analysis in hematopietic stem cell transplanation. In: Atkinson K (eds). *Clinical Bone Marrow and Blood Stem Cell Transplantation*, 2nd edn. Cambridge University Press, 2000, pp 1415–1423.
8 Szydlo R, Goldman JM, Klein JP *et al*. Results of allogeneic bone marrow transplants for leukemia using donors other than HLA-identical siblings. *J Clin Oncol* 1997; **15**: 1767–1777.
9 Gooley TA, Leisenring W, Crowley J, Storer B. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med* 1999; **18**: 695–706.
10 Pepe MS, Longton G, Pettinger M *et al*. Summarizing data on survival, relapse, and chronic graft-versus-host disease after bone marrow transplantation: motivation for and description of new methods. *Br J Haematol* 1993; **83**: 602–607.
11 Pepe MS, Mori M. Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med* 1993, **12**: 737–751.
12 Altman DG. *Practical Statistics for Medical Research*, section 13.4.6. Chapman Hall: London, 1991.
13 Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965; **52**: 203–223.
14 Tarone RE, Ware JH. On distribution-free tests for equality for survival distributions. *Biometrika* 1977; **64**: 156–160.
15 Fleming TR, Harrington DP. A class of hypothesis tests for one and two samples of censored survival data. *Comm Stat* 1981; **10**: 763–794.
16 Kolb HJ, Mittermuller J, Clemm CH *et al*. Donor leukocyte transfusions for treatment of recurrent chronic myelogenous leukemia in marrow transplant patients. *Blood* 1990; **76**: 2462–2465.
17 Collins RH, Shpilberg O, Drobyski WR *et al*. Donor leukocyte infusions in 140 patients with relapsed malignancy after allogeneic bone marrow transplantation. *J Clin Oncol* 1997; **15**: 433–444.
18 Kolb HJ, Schattenberf A, Goldman JM *et al*. Graft-versus leukemia effect of donor lymphocyte transfusions in marrow grafted patients. *Blood* 1995; **86**: 2041–2050.
19 Sehn LH, Alyea EP, Weller E *et al*. Comparative outcomes of T cell-depleted and non-T cell-depleted allogeneic bone marrow transplantation for chronic myelogenous leukemia: impact of donor lymphocyte infusion. *J Clin Oncol*.1999; **17**: 561–568.
20 Craddock C, Szydlo, R, Olavarria, E *et al*. Leukemia-free survival after allogeneic transplantation for chronic myeloid leukemia: effect of reclassifying responders to donor lymphocyte infusions as 'currently free of leukemia'. *Blood* 1997; **90** (Suppl): 327b.
21 Klein JP, Szydlo RM, Craddock C, Goldman JM. Estimation of current leukemia-free survival following donor lymphocyte infusion therapy for patients with leukemia who relapse after allografting: application of a multistate model. *Stat Med* 2000; **19**: 3005–3016.
22 Klein JP, Keiding N, Shu Y *et al*. 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.

Appendix: Glossary of statistical terms

Bonferroni correction This correction is used when multiple hypothesis tests are made on the same set of data. This correction provides insurance that the likelihood of falsely rejecting any one of k hypotheses is at most some fixed value, α, when each null hypothesis is true. The Bonferroni correction is made by using an α/k level cut-off for each of the k tests performed on the data. This correction makes the experiment-wise error approximately equal to α. The experiement-wise error rate is the probability of making an incorrect decision for any of the hypotheses when all the null hypotheses are true.

Censored data A patient's event time is censored if we have only partial information on when the event would occur. Most common is right-censoring where we only know that the patient's event time has yet to occur. Other types of censoring are left-censoring, where all we know is that the event occurred prior to some time but not exactly when, and interval-censoring, where all we know is that the patient's event time occurred in some window of time.

Chi-square test A test used to compare k treatment groups when all we know for each patient is yes or no has the event occurred or not. If O_{ij} is the observed number of patients in the jth group with a response i = yes or no and E_{ij} is the number of patients we would expect in this category if there were no difference in treatments, then the chi-square statistic is the sum of $(O_{ij} - E_{ij})^2/E_{ij}$. In other words the chi-square test is the sum of the squared differences of the observed and expected counts scaled by the expected counts. Statistic is compared to a chi-square table with $k - 1$ degrees of freedom.

Competing risks Two or more events where occurrence of one of the events precludes occurrence of any other of the events.

Confidence band (for the survival function) A 95% confidence band is an upper and lower curve with the property that we are 95% confident that the true survival function is between these two curves for any time-point. Confidence bands tend to be wider than confidence intervals (see below) since we are making a promise that the entire curve is in the band, not simply the value at one point in time.

Confidence interval A 95% percent confidence interval for the survival time at some fixed point in time is an interval with the property that we are 95% sure that the true value of the survival function is in this interval. The inference is to the survival function at a single point in time.

Covariates These are independent variable or risk factors known at the time of transplant which may influence outcome.

Cox model A model that relates covariates or risk factors to outcome. The model, first proposed by Sir David Cox,[1] assumes that the hazard rate for an individual with a set of covariates is the product of a baseline hazard rate and a parametric function of the risk factors.

Crude hazard rate In a competing risk model, this is the rate at which a patient who has yet to experience any of the competing risks will experience a particular competing risk.*Cumulative incidence* In a competing risks framework, the chance a patient will have experienced a particular competing risk prior to time *t*. It is probability in the real world where a patient may experience any of the competing risks.

Current leukemia-free survival The chance a patient is alive in remission (first or subsequent) at some time after transplant. Of particular interest when patients may achieve a remission after a post-transplant relapse.

Disease-free survival The chance a patient is alive and disease-free at a given point in time after transplant.

Fleming and Herrington weighted log rank test A weighted log rank test of the hypothesis of no difference between groups with weight $S(t)^p(1 - S(t))^q$, where S(t) is the pooled Kaplan–Meier estimator at time *t*. When $p = 1$ and $q = 0$ more weight is given to early departures, while when $q = 0$ and $p = 1$ more weight is given to late departure between the groups.

Gehan's test A weighted log rank test of the hypothesis of no difference between groups, where the weight is the number at risk at each event time. The test gives more weight to early differences in outcomes between groups.

Hazard rate The rate at which an event is occurring. The hazard rate of death is approximately the chance of dying in the next moment in time among current survivors. The hazard rate is the slope of the complement of the survival function.

Kaplan–Meier estimate An estimate of the survival function or disease-free survival function suggested by Kaplan and Meier.[3] This estimator is also known as an actuarial estimate of survival. It is not appropriate for competing risks data.

Left truncation Left truncation occurs when subjects enter a study at a random time upon occurrence of some event (not necessarily the origin for the event of interest) and are followed from this delayed entry time until the event occurs or until the subject is censored. Note this is distinct from censored data where some information is observed for every patient

Log rank test A weighted log rank test with weight of 1. This test gives equal weight to differences in groups at any point in time.

Nonparametric procedure A statistical method which makes no assumptions about a model for the data. Nonparametric tests are also known as distribution free tests. Two sample nonparametric tests tend to compare the complete distribution of the data in each sample, not some measure of the distribution in each sample. In the complete data setting these are typically tests based on ranks and simple counts.

One way analysis of variance Analysis of variance, or ANOVA, is the classical test of the hypothesis of no difference between the means of two or more treatment groups based on continuous responses from each patient.

P value The smallest type I error at which the null hypothesis will be rejected. This is also known as the observed significance level.

Parametric procedures These are procedures which assume that the data came from a family of statistical distributions indexed by some parameter. They require an assumption of a model for the data with some unknown parameter. Tests are about some summary population measure (a parameter) such as the mean or variance. Typical examples are the analysis of variance or the *t*-test for uncensored data which assumes that data are from a normal distribution. These tests can be misleading if the assumed model is wrong, but they are usually more powerful when the model is correct.

Power of the test The chance of rejecting the null hypothesis. When the null hypothesis is true this is the chance of a type I error (the significance level). When the null hypothesis is false it is the chance of making a correct decision.

Quadratic form A mathematical construct for computing test statistics. If **A** is the vector $([p_1 - p_2], [p_1 - p_3, \ldots, [p_{k-1} - p_k])$ where p_j is the estimate in the jth group and **V** is the estimated matrix of variance-covariances of **A**, then a quadratic form is $\mathbf{AV}^{-1}\mathbf{A}^t$.

Risk set The collection of all subjects who could have experienced an event at time *t*.

Significance level The chance of falsely rejecting the null hypothesis. It is the type I error probability.

Semi-parametric model A model or procedure where the data are modeled by a combination of an arbitrary function and a specific parametric model. The Cox model is a semi-parametric model since we model the hazard rate for an individual by a general baseline function (the non-parametric part) times a specific parametric function of the covariates.

Standard deviation The square root of the variance. It is a measure of dispersion.

Standard error The standard deviation of an estimator. It is a measure of how precise an estimator is.

Survival function The chance a patient is alive at time *t*.

Tarone and Ware's weighted log rank test A weighted log rank test where the weight is the square root of the number at risk at each event time. The test gives more weight to early differences in outcomes between groups

Time-dependent covariates Prognostic or risk facts whose value changes over time. Examples would be weekly white blood counts, occurrence of GVHD, etc.

Type I error The error of rejecting a null hypothesis when it is in fact true.

Type II error The error of not rejecting a null hypothesis when it is false.

Weighted log rank test A test comparing k groups which is a weighted sum of the squared difference between the observed number of deaths and the expected number of deaths in each group at each event time.