Department of medical statistics

Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls

Stuart J Pocock, Tim C Clayton, Douglas G Altman

Survival plots of time-to-event data are a key component for reporting results of many clinical trials (and cohort studies). However, mistakes and distortions often arise in the display and interpretation of survival plots. This article aims to highlight such pitfalls and provide recommendations for future practice. Findings are illustrated by topical examples and also based on a survey of recent clinical trial publications in four major journals. Specific issues are: should plots go up or down (we recommend up), how far in time to extend the plot, showing the extent of follow-up, displaying statistical uncertainty by including SEs or CIS, and exercising caution when interpreting the shape of plots and the time-pattern of treatment difference.

In many clinical trials, the primary outcome for comparison of treatments is the time to occurrence of a disease-related event. The most widely adopted method of displaying such results is by means of Kaplan-Meier survival plots, which show the proportion of patients who experience (or do not experience) the event by time since randomisation. The event itself could be death (hence the term "survival plot" is used loosely), but is often time to a non-fatal event (eg, disease recurrence in cancer) and can sometimes be a favourable outcome such as discharge from hospital. Combined endpoints are used increasingly in clinical trials (eg, death, acute myocardial infarction, or cardiac arrest), and in such cases, the survival plot shows the time to the first event.

The statistical methods for producing survival plots and for calculating p values, estimates of treatment effects, and associated CIs are all well documented.[1–3] However, the display and interpretation of survival plots are prey to several potential distortions and deceptions that can make the right message difficult to work out, as reported in a previous survey of survival analyses in cancer trials.[4] In this article, we concentrate on treatment comparisons in clinical trials, although many of the same problems apply to survival plots in general. Our aim is to reveal some of the more common pitfalls and to give some guidelines to authors, journal editors, and readers on what constitutes desirable statistical practice.

As a practical basis for our concerns and conclusions, we identified all 35 clinical trials with survival plots that were published in four general medical journals during July to October, 1999 (19 in *The Lancet*, ten in the *New England Journal of Medicine*, four in the *British Medical Journal*, and two in the *Journal of the American Medical Association*). These trials constituted 41% of the 86 individually randomised parallel-group trials published in the four journals.

Lancet 2002; **359:** 1686–89

Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK (Prof S J Pocock PhD, T Clayton MSc)**; Centre for Statistics in Medicine, Oxford, UK** (Prof D G Altman DSc)

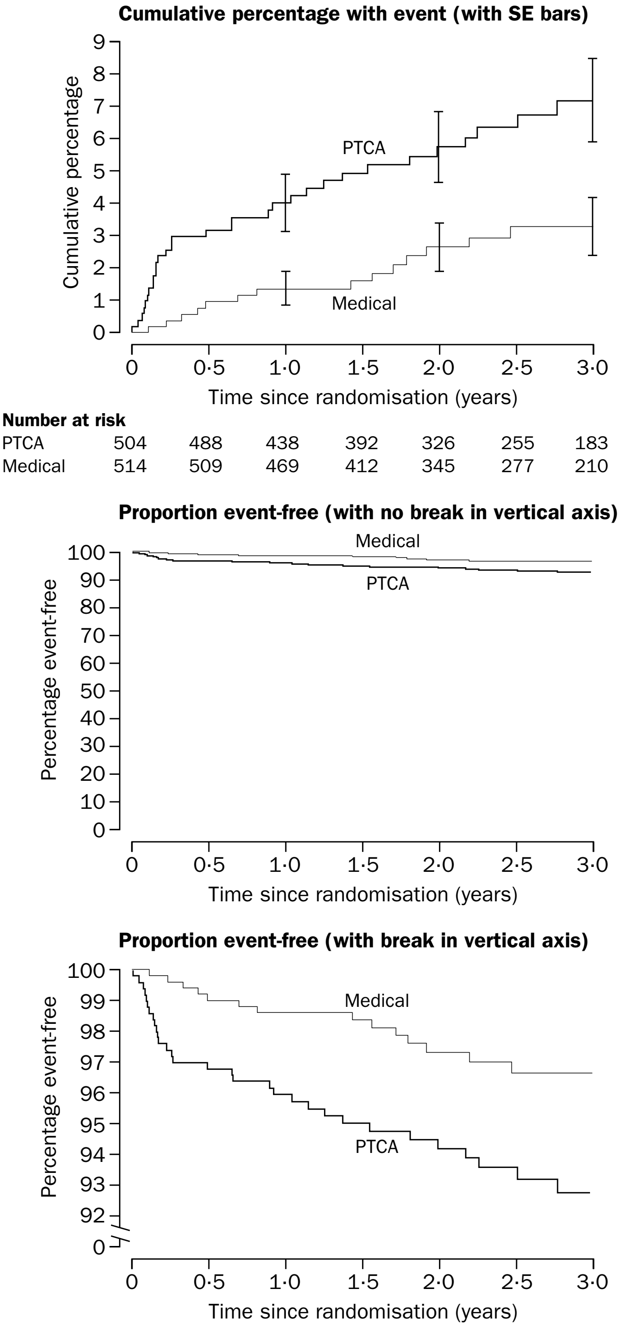
Correspondence to: Prof Stuart J Pocock
(e-mail: stuart.pocock@lshtm.ac.uk)

Should plots go up or down?

A survival plot going down displays the proportion of patients free of the event (which of course declines over time), whereas a plot going up shows the cumulative proportion experiencing the event by time. In principle, both contain the same information, but the visual perceptions with regard to comparison of treatment groups can be quite different.

For instance, figure 1 shows three ways of displaying the same data on time to non-fatal myocardial infarction or death in the RITA-2 trial.[5] The first plot, going up, indicates clearly the excess of events in the group randomised to percutaneous transluminal coronary angioplasty (PTCA) compared with the group continuing on medical treatment. This plot has the same style as in the trial's publication,[5] which also gave the numbers and percentages of patients with myocardial infarction or death: 32 of 504 (6·3%) and 17 of 514 (3·3%) for the PTCA and medical treatment groups, respectively (p=0·02). The second plot, going down and using the whole vertical axis from 0 to 100%, makes the difference look much less pronounced (the corresponding proportions event-free being 93·7% and 96·7%, respectively) and mainly emphasises that most patients did not experience the event. The third plot, going down but with a break in the vertical axis seems to fill the space more informatively, but relies on the reader recognising the break in scale: if they do not, the impression is left that PTCA is harmful to a large proportion of patients. Hence having such a break in the scale is not a good style to adopt.

In practice, only one of these options can be displayed in a trial report. We recommend the first option—the plot going up—as the most reliably informative, especially if the event rate is lower than, say, 30%. To maximise the clarity of information, the highest value on the vertical axis should be a round number slightly greater than the highest value represented by the steepest curve—ie, 9% in figure 1. Some might argue that the full scale (0–100%) should always be included, but this inhibits the ability to discriminate between treatments. For instance, the ELITE 2 trial[6] included such a plot, which helped to hide the apparent survival inferiority of losartan compared with captopril. Admittedly, the treatment difference was not statistically significant, but any claim of potential equivalence was perhaps falsely magnified by the choice of survival plot going down over the full 100%



Number at risk

PTCA	504	488	438	392	326	255	183
Medical	514	509	469	412	345	277	210

Figure 1: **Time to non-fatal myocardial infarction or death in RITA-2 trial: three ways to display same data**

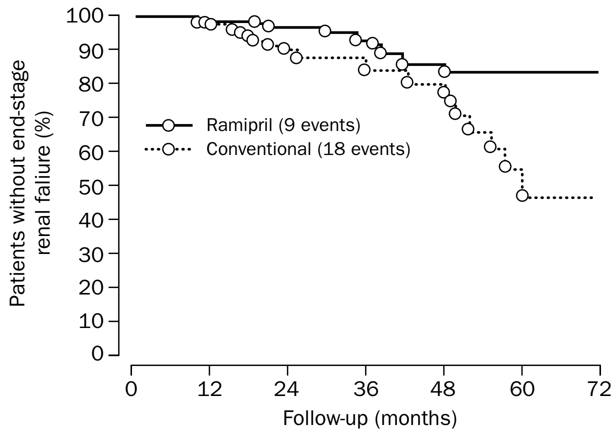


Figure 2: **Kaplan-Meier estimation of renal survival among patients on ramipril or conventional treatment**
Relative risk 2·72 (95% CI 1·22–6·08), p=0·01.

scale.[7] The important survival superiority of pravastatin over placebo in the LIPID trial[8] was hard to discern because of this same injudicious choice of survival plot going down over the full 100% scale, since death rates in all groups were, in fact, less than 10% after 5 years. Incidentally, the investigators claim that this choice was introduced by the journal, not the authors themselves. Such plots going down are useful only for trials in which the event rate is high, such as those in cancers with poor prognosis. For instance, for a neuroblastoma trial,[9] the same style of survival plot was perfectly clear, since the median survival was less than 2 years in a study with follow-up over 5 years for those still alive.

The applicability of trial findings should not rely on relative treatment differences alone (eg, proportional reduction in mortality), but must also include absolute treatment differences (eg, number needed to treat per life saved[10,11]). Provision of both survival plots would perhaps be ideal, one going up to reveal the detail and the relative treatment difference, and one going down to clarify the small absolute risk and hence small absolute difference in treatments. In trial reports for which space is at less of a premium and in regulatory submissions, such an approach is to be encouraged for the key outcomes, but it is unrealistic for journal publications.

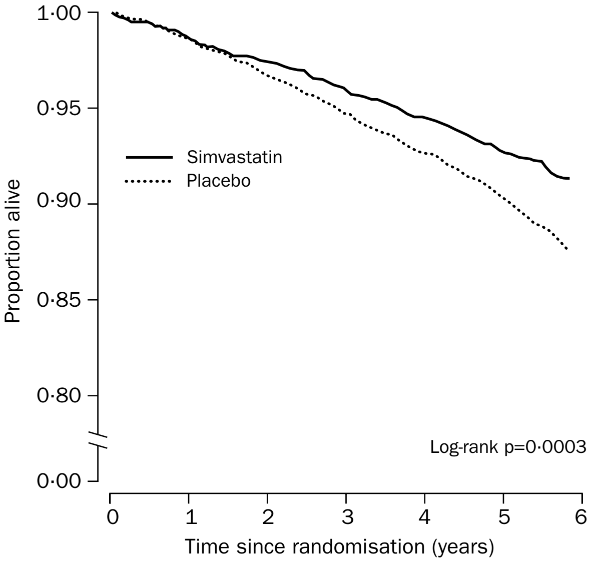
In the 35 trials we surveyed, 12 had plots going up, 15 had plots going down all the way to zero, and eight had plots going down but with a break in scale. This disparity in approach is undesirable.

How far in time to extend the plot?

Follow-up times in any one trial can vary substantially because patients are usually recruited over a long period, and some patients can be lost to follow-up. Length of follow-up is taken into account in the Kaplan-Meier life-table method[1–3] for estimating the proportion of patients who experience an event by time since randomisation. Technically, any survival plot can be extended right through to the longest follow-up time, and five trials we surveyed did just that. However, this extension is not good statistical practice, since for any such plot the eye is drawn to the right (ie, where the plot finishes), which is where there is least information and greatest uncertainty. In small trials, much of the right-hand part of the plot can depict just a few patients.

For instance, figure 2 is a reproduction of the plot of time to end-stage renal failure in a trial comparing ramipril with conventional treatment.[12] The visual impression is that treatments are similar up to 48 months, but thereafter the conventional group develops a striking excess of end-stage renal failures, reaching an estimated 50% failure, by 60 months. However, the median follow-up was 31 months and only 25% of patients assigned conventional treatment reached 48 months' follow-up. The number reaching 60 months is not stated but must be very few. Thus, for both treatment groups, there are inadequate data to estimate reliably the failure rates beyond 48 months' follow-up.

In general, we recommend that survival plots be halted once the proportion of patients free of an event, but still in follow-up, becomes unduly small. In our experience, this view is not universally held, but we hope that our recommendation is a good basis for debate.



Number at risk

Simvastatin	2221	2193	2160	2131	2097	2060	113
Placebo	2223	2193	2152	2103	2059	2011	115

Figure 3: **Kaplan-Meier curves for all-cause mortality in 4S trial**

What constitutes "unduly small" is open to debate and depends on the context. It will often be reasonable to curtail the plot when only around 10–20% are still in follow-up. For example, suppose in a trial of 500 patients, 100 had the event of interest by 2 years of follow-up, but of the remaining 400 patients, only 80 (20%) were still in follow-up beyond 2 years. In this case, restriction of the plot to 2 years' follow-up might be sensible. Such a restriction is just for the plot; all events should be retained in analysis (eg, nine *vs* 18 events in the ramipril trial should remain the basis for the statistical inference given in the legend of figure 2). In this example, the authors' dilemma is clear, since all the "action" happens beyond 48 months. However, were the later follow-up to be included in the plot, it should include a note highlighting the small number of patients on which the data were based. These problems do not arise for trials with an intended fixed length of follow-up (usually quite short), as was the case for 21 of the trials we surveyed.

Showing the extent of follow-up

So, readers need to be informed about the extent of follow-up, and stating the median follow-up time is often useful. Another helpful device is to display the numbers of patients event-free and still in follow-up in each treatment group at relevant time points, as shown in figures 1 and 3. These numbers at risk of the event convey to the reader the increasing unreliability of estimates as time gets further from randomisation; most trials we surveyed included this information. The numbers on the time axis of the published 4S trial plot[13] reproduced in figure 3 show a case for not extending the graph to 6 years. Since only a small minority of patients reached 6 years' follow-up, the apparent extra boost in treatment difference in that last year is less reliably estimated. Incidentally, plots going downwards with an axis break like figure 3 make focusing on the main finding harder. We needed a ruler and calculator to work out that mortality rates after 5 years were 7·4% on simvastatin and 9·7% on placebo.

Displaying statistical uncertainty

Most outcome results of clinical trials include measures of statistical uncertainty—eg, either SEs or CIs—for each treatment group, or a CI for the comparison of groups. However, survival plots often fail to include such measures. Hence the visual impression of any treatment differences, and how they vary over time, can look much more convincing than is really the case, especially if the clinical trial has few outcome events.

For any time since randomisation, the SE (or 95% CI) for the estimated proportion of patients with (or without) the event can be calculated.[2,3] In principle, such error bands could be displayed at all time points for each treatment group, but displaying the SE or 95% CI at a few regularly spaced time points on the plot for each treatment group is clearer. For instance, figure 1 (top panel) shows the SE bars for the estimated event rate for each treatment at 1, 2, and 3 years' follow-up. As is common in such plots, the smaller numbers of patients in follow-up at later time points is reflected in the increasing SE over time.

Although these SEs display each plot's uncertainty, they do not directly display the uncertainty of the treatment difference, which is usually of primary interest. In fact, the SE of the treatment difference in event rates is equal to the square root of the sum of the two squared SEs, but there is no conventionally accepted style (nor any easy way) of displaying this on a survival plot. One simple rule of thumb is that if the treatment difference at a particular time is less than the sum of the two plotted SEs (ie, if the plotted SEs overlap), the difference is well within the bounds of random chance. If the difference is more than twice the sum of the SEs (ie, the 95% CIs do not overlap) it is highly significant. Whether SEs or 95% CIs should be plotted is open to debate, but authors should always make clear which is being used.

One problem here is the focus on the difference between treatments at particular arbitrary time points. The overall evidence of a treatment difference is usually given by the estimated hazard ratio (sometimes called relative risk) and its 95% CI,[14] and by a log-rank test of significance,[2,3] as shown in the legend of figure 2. Thus, an alternative to plotting SEs is to present overall treatment comparisons and their uncertainty on the survival plot or its legend. Most authors do neither, leaving any comment on statistical uncertainty to the text only. In fact, only one of the 35 trial reports we surveyed included CIs at regular time points on the survival plot, five plots included the hazard ratio and its CI, and 16 plots incorporated the log-rank p value.

Summary of recommendations

- Survival plots are best presented going upwards, to maximise detail without needing a break in the scale
- Plots should only be extended through the period of follow-up achieved by a reasonable proportion of participants
- The extent of follow-up should be explained—eg, by listing at regular intervals under the time axis the number still at risk in each treatment group
- Plots should include some measure of statistical uncertainty, otherwise any visual signs of treatment differences might look more convincing than they really are. Either SEs or CIs should be displayed at regular time points, or an overall estimate of treatment difference (eg, relative risk) with its 95% CI should be given
- Authors and readers should be cautious in interpreting the shape of survival plots. The lack of follow-up and poorer estimation to the right-hand end, the lack of any prespecified hypothesis, and the lack of statistical power to explore subtleties of treatment difference other than the overall comparison should be recognised

18 plots included none of the above. We recommend that future authors include in each survival plot some indication of statistical uncertainty (panel).

Interpreting the shape of survival plots

The easiest patterns to interpret are those that show no apparent difference between treatments or when there is a steady divergence between treatments over time. However, in many instances, more complex patterns seem to exist: the treatment difference might look greater early on (figure 1), the divergence between treatments might start later on (figures 2 and 3), or the survival curves might cross. Such putative treatment–time interactions need cautious interpretation since there are rarely sufficient data to consolidate their true existence.

For instance, in the ramipril trial (figure 2), most of ramipril's benefit seems to have occurred late: nine ramipril failures versus 11 conventional treatment failures before 48 months, compared with zero failures versus seven failures, respectively, after 48 months. However, the strength of evidence for this effect is limited, since the number of failures is small, the statistical test for treatment–time interaction is of borderline significance, and such a post-hoc (data-driven) analysis is disputable. So, the overall conclusion needs to rest on events during the total follow-up rather than after any specific time point.

Even for the much larger 4S trial (figure 3), caution is required in interpreting the visual impression that the treatment effect does not occur until after 18 months' follow-up. There seem to have been 55 deaths in each group in the first 18 months, and a striking treatment difference thereafter, with 201 versus 127 deaths favouring simvastatin.[13] A test for treatment–time interaction (ie, of whether the hazard ratio is different before and after 18 months) is significant ($p=0{\cdot}03$), but its validity can be questioned because the 18-month time-split for the data has been selected post hoc after seeing the survival plot. Thus, even in such a large trial, to expect reliable estimation of when a treatment effect first begins is unrealistic.[15] Indeed, recent evidence from the Heart Protection Study indicates that there is an observable treatment difference in survival even in early follow-up, which becomes more rapidly divergent beyond 2 years (www.hpsinfo.org).

References

1 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; **53:** 457–81.

2 Collett D. Modelling survival data in medical research, section 2.1. London: Chapman and Hall, 1994.

3 Altman DG. Practical statistics for medical research, chapter 13. London: Chapman and Hall, 1991.

4 Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995; **72:** 511–18.

5 RITA-2 Trial Participants. Coronary angioplasty versus medical therapy for angina: the second Randomised Intervention Treatment of Angina (RITA-2) trial. *Lancet* 1997; **350:** 461–68.

6 Pitt B, Poole-Wilson A, Segal R, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial—the Losartan Heart Failure Survival Study ELITE II. *Lancet* 2000; **355:** 1582–87.

7 Hall A. Comparison of losartan and captopril in ELITE II. *Lancet* 2000; **356:** 851.

8 Tonkin AM, Colquhoun D, Emberson J, et al. Effects of pravastatin in 3260 patients with unstable angina: results from the LIPID study. *Lancet* 2000; **356:** 1871–75.

9 Matthay KM, Villablanca JG, Seeger RC, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-*cis*-retinoic acid. *N Engl J Med* 1999; **341:** 1165–73.

10 Altman DG, Anderson PK. Calculating the number needed to treat for trials where the outcome is time to an event. *BMJ* 1999; **319:** 1492–95.

11 Lubsen J, Hoes A, Grobbee D. Implications of trial results: the potentially misleading notions of number needed to treat and average duration of life gained. *Lancet* 2000; **356:** 1757–59.

12 Ruggenenti P, Perna A, Gherardi G, et al. Renoprotective properties of ACE-inhibition in non-diabetic nephropathies with non-nephrotic proteinuria. *Lancet* 2000; **354:** 359–64.

13 Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994; **344:** 1383–89.

14 Altman DG, Machin D, Bryant TN, Gardner MJ, eds. Statistics with confidence, chapter 9. London: BMJ Publishing, 2000.

15 Boutitie F, Gueyffier F, Pocock SJ, Boissel J-P. Assessing treatment-time interaction in clinical trials with time to event data: a meta-analysis of hypertension trials. *Stat Med* 1998; **17:** 2883–903.