Data maturity and follow-up in time-to-event analyses

Val Gebski, MStat, Valérie Garès, PhD, Emma Gibbs, MSc, Karen Byth, PhD

National Health and Medical Research Council Clinical Trials Centre, University of Sydney, Sydney 2006, Australia

Correspondence to: Professor Val Gebski, NHMRC Clinical Trials Centre, Locked Bag 77, Camperdown NSW 1450, Australia

Email: val@ctc.usyd.edu.au; Tel +61 2 9562 5000.

Abstract

We propose methods to determine the minimum number of subjects remaining at risk after which Kaplan-Meier survival plots for time-to-event outcomes should be curtailed, as, once the number remaining at risk drops below this minimum, the survival estimates are no longer meaningful in the context of the investigation. The size of the decrease of the Kaplan-Meier survival estimate $S(t)$ at time t if one extra event should occur is considered in two ways. In the first approach, the investigator sets a maximum acceptable absolute decrease in $S(t)$ should one extra event occur. In the second, a minimum acceptable number of subjects still at risk is calculated by comparing the size of the decrease in $S(t)$ if an extra event should occur to the variability of the survival estimate had all subjects been followed to that time (confidence interval approach). We recommend calculating both limits for the number still at risk and then making an informed choice in the context of the particular investigation. We explore further how the amount of information actually available can assist in considering issues of data maturity for studies whose outcome of interest is a survival percentage at a particular time point.

We illustrate the approaches with a number of published studies having differing sample sizes and censoring issues. In particular, one study was the subject of some controversy regarding how far in time the Kaplan-Meier plot should be extended. The proposed methods allow for limits to be calculated simply using the output provided by most statistical packages.

Introduction and motivation

For studies with time-to-event outcomes, the associated survival estimates need to account for censoring which occurs when individuals do not experience the event of interest during the period of their follow-up. The Kaplan-Meier method produces survival estimates $S(t)$ at time t which are then displayed graphically as a Kaplan-Meier plot (survival curve)[1, 2]. This curve is a step function, decreasing each time an event is observed. It is used extensively in medicine and health sciences where time-to-event outcomes are common, and to illustrate data describing disease history available from clinical registries, to monitor the failure of medical devices, in areas of health economics and so on. A common application of the Kaplan-Meier approach is to obtain estimates and associated confidence intervals (CIs) of the probability of remaining event-free at specific time points. These estimates (a) help inform clinical practice and choice of therapy; (b) assist in evaluating new therapies; and (c) facilitate the evaluation of public health policies in relation to resources or directions required to reduce disease burdens in populations of interest.

The uncertainty of a survival estimate increases as the number in the sample remaining at risk decreases over time. Survival estimates are often presented over the full duration of the follow-up period, when more thought could be given to the value of the information displayed[3-5], a point highlighted by Carter [6]. Therefore, even though the estimates *can* be calculated, the question arises as to whether they *should* be included in the Kaplan-Meier plot? In other words, how far in time should the Kaplan-Meier plot be extended? This question is directly related to the number of events which have been observed, and the extent of the study follow-up (data maturity).

In prospective clinical trials, data maturity is one of the design parameters used to determine sample size. Ideally such studies do not report their results until the predetermined follow-up time is reached. The issue of data maturity is crucial in prospective and retrospective observational studies such as disease registries and epidemiological investigations, especially when the focus is the survival estimate at a clinically important time point. When long-term follow-up is of interest, and the evaluation time has not been pre-specified, a Kaplan-Meier plot should be curtailed before the number of subjects remaining at risk is too small to meaningfully interpret the survival

estimates. Guidelines regarding how far in time to extend the Kaplan-Meier plot would help to minimise interpretations which may be misleading. Pocock *et al*[7], suggested "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…. It will often be reasonable to curtail the plot when only around 10–20% are still in follow-up." However, this guideline is problematic if the original study size is large: for example, 10-20% of a sample of 500 subjects corresponds to 50-100 subjects still at risk. Clark *et al*'s measure of data completeness[8] provides an index of the actual follow-up, which may be reduced by drop-out and censoring, relative to the expected follow-up if all subjects were accounted for to the end of the study. This index helps guide the interpretation of trial results. However in many observational cohorts, and especially clinical registries, the expected follow-up time for individuals in the cohort is unknown. To address some of the drawbacks relating to uncertainty and censoring, methods have been proposed to adjust the variance of the survival estimate to account for the reduction in the number at risk due to censoring resulting in modified variance estimators[9-14].

Apart from the issue of precision of the survival estimate when only a small number of subject are still at risk, the interpretation of the survival estimate may suffer from representiveness bias and compromise generalizability. For example, the trial of radioembolization for metastatic liver cancer in a high-risk patient cohort[15] reported one patient (receiving the intervention) remaining disease free for over 8 years, while for all the other patients the disease-free survival (DFS) was < 5.5 years. Such a long DFS was considered extraordinary, but in subsequent similar studies with over 1000 (lower risk) patients and more modern chemotherapy[16] a DFS of this magnitude could not be achieved. This individual was clearly atypical of this population and long term DFS rates based on this study would be misleading.

A related question of data maturity is 'For how long should subjects be followed once recruitment has been completed in studies whose primary outcome is the proportion of subjects remaining event-free at a particular time?' This question often arises in studies of surgical techniques, organ transplantation, medical devices, paediatric therapies and radiotherapy where comparisons of 5

---

or 10-year survival rates are of more clinical interest than hazard ratios[17-19]. Consider, a study of adjuvant therapy after renal transplantation which has just completed accrual with the outcome being 5-year rejection-free survival. What proportion of patients should have completed 5-years follow-up, at the time of the analysis? Should the analysis wait until the last patient has reached 5-year follow-up or can the data be analysed sooner, and if so, when? Achieving complete follow-up may be inefficient and costly especially if only a small number of subjects were recruited in, say, the past 18 months. Prolonging follow-up to ensure the study achieves a 100% follow-up target may be impracticable or may have only a marginal impact on study results compared to those from an assessment at some earlier time.

How far in time to extend the Kaplan-Meier survival plot

Consider a sample of N subjects followed up over time until the event of interest occurs. Let $S(t)$ denote the Kaplan-Meier event-free survival probability estimate at time t when $n(t)$ subjects remain at risk. At time $t=0$, $n(0)=N$, $S(0)=1$ and subsequently decreases as events occur. If one extra event had occurred immediately after time t, then the *decrease in the estimated percentage of subjects event-free* would be $\Delta(t)$ where

$$\Delta(t) = 100S(t)/n(t) \qquad \text{is defined for } n(t) \geq 1.$$

We will refer to $\Delta(t)=100S(t)/n(t)$ as the *sensitivity index* of the survival estimate at time t.

If the estimated event-free survival probability at time t is high and few subjects remain at risk, then the sensitivity index will be large, indicating the potential for a sharp drop in the Kaplan-Meier plot due to a single extra event. This is illustrated in figure 1 using data from the I-ELCAP study[20], showing the survival outcomes of 484 asymptomatic patients with early stage lung cancer diagnosed through a screening program. The 10-year %survival estimate of 80% features prominently throughout the report. However, the sensitivity index at 10-years, shown as the thick dashed line, is 40% which is based on just two patients remaining at risk. This indicates a high sensitivity of the 10-year survival estimate to a single extra event which would drop the estimated 80% survival to 40%. Dashed/dotted arrows indicate the 10% and 20% number remaining at risk limits (at 60 and 72 months respectively) for extending the Kaplan-Meier plot suggested by

Pocock[7]. The lower one-sided 95% confidence band developed by Fay, Brittain and Proschan[11] to account for increasing uncertainty due to the decreasing number at risk arising from censored follow-up is also displayed.

// Figure 1 Kaplan-Meier plot for International Early Lung Cancer Investigators (I-ELCAP)

Study//

We propose two approaches based on the sensitivity index $\Delta(t)$ for deciding when to curtail a Kaplan-Meier plot. Both produce a minimum number remaining at risk required to satisfy a particular criterion. We suggest using these values to decide on a suitable curtailment point in the context of the investigation.

Criterion 1. Pre-defined sensitivity index threshold Δ^*

The first criterion sets a maximum acceptable decrease Δ^* in the estimated percentage of subjects event-free should one extra event occur. That is, $\Delta(t) = 100S(t)/n(t) < \Delta^*$, for all points displayed on the Kaplan-Meier plot. Therefore, for all t up to and including the curtailment time,

$$n(t) > 100S(t)/\Delta^*. \qquad (1)$$

This criterion is particularly helpful for studies exhibiting a high degree of 'early' censoring, (patients at risk not having sufficient follow-up over the study duration) as in the I-ELCAP study. When considering clinical practice or developing guidelines, Δ^* should be small, <1% say (for example when using a clinical registry with large patient numbers). For other clinical studies (with 'small' or 'moderate' sample sizes) larger threshold values (2.5% or 5%) might be considered acceptable.

Statistical packages routinely provide details of $n(t)$ and $S(t)$ at all censoring or event times t. The survival plot can be extended to t and satisfy Criterion 1 provided $n(t) > 100S(t)/\Delta^*$. When the survival curve is close to the x axis, this issue may not be so important[3, 21].

Criterion 2. $\Delta(t)$ < width of one-sided 95% CI for %survival based on 'full information'

Criterion 1 however does not consider the uncertainty present in a survival estimate at the time point of interest, *t*. If there has been no censoring up to or including time *t*, then *all N* subjects enrolled in a study will have been followed up to or past time *t*. In this optimal situation of *full information at time t*, the standard error *SE(t)* of the estimate *S(t)* is smaller than if some censoring had occurred prior to time *t*. When there is full information at time *t*, *SE(t)* is simply the standard error of a binomial proportion *S(t)* with sample size *N* (the number of subjects entering the study), that is $SE(t) = \sqrt{\{S(t)[1-S(t)]/N\}}$. Furthermore, the lower boundary of the one-sided 95% confidence interval (CI) for the estimated percentage of subjects event-free in the case of full information at time *t* is $100*[S(t)-1.645*\sqrt{\{S(t)[1-S(t)]/N\}}]$ where 1.645 is the upper 5% quantile (often written $z_{1-\alpha}$) of the standard normal distribution. We refer to this boundary as the *full information one-sided 95% confidence boundary of %survival.*

Our second criterion requires the sensitivity index *Δ(t)* to be no larger than the width of the full information one-sided 95%CI for the %survival at *t*, that is

$$\Delta(t) = 100*S(t)/n(t) \le 100*1.645*\sqrt{\{S(t)[1-S(t)]/N\}}, \text{ which gives}$$

$$n(t) \ge \frac{1}{1.645}\sqrt{\frac{NS(t)}{1-S(t)}}. \tag{2}$$

This criterion implies that one extra event observed just after time *t* would not decrease the estimated %survival to below its *full information one-sided 95% confidence boundary at time t.* We interpret this as evidence that enough subjects remain at risk at time *t* for the meaningful interpretation of the Kaplan-Meier plot.

The relationship between the study size *N* and the minimum number at risk required to satisfy (2) is shown in figure 2 for *S(t)* =0.1,0.2, …, 0.95. The different levels of *S(t)* reflect the different amounts of censoring present. Again, the quantities *N* and *S(t)* required in (2) are routinely produced by statistical packages. If a less conservative bound is desired for Criterion 2, the full information lower one-sided 97.5% CI can be used instead. In this case the divisor of 1.645 in (2) is replaced by 1.96.

//Figure 2 Minimum number of patients at risk satisfying Criterion 2 for different sample sizes *N* and selected values of the Kaplan-Meier curve *S(t)* from 0.1 to 0.95//

Application to published studies

Figure 1 (from reconstructed data[12]) shows the survival curve and the corresponding number at risk for 484 asymptomatic lung cancer patients. Of particular interest was the 120-month survival rate but quoting this rate when there were only 2 patients remaining at risk has received some discussion[6, 12]. Details of our approach are provided in table 1, which examines when the Kaplan-Meier plot should be curtailed.

//   Table 1: Number at risk of an event in the I-ELCAP lung study //

In order to satisfy Criterion 2, the minimum number at risk at 84 months (shown by the solid vertical line in figure 1) is $n \geq 27 = \{\sqrt{484}*[0.8/0.2]\}/1.645$. This criterion is not satisfied for times after 84 months and extending the plot beyond this time would not be recommended. Additionally, the sensitivity index of the estimate at 84 months is 2.72% which, for this moderate sized study, appears reasonable. The 10-20% rule suggested by Pocock et al. would suggest curtailment at 59 or 72 months (indicated by the vertical dashed lines in figure 1) , when 96 or 48 subjects are still at risk.

A second example, the Continuous Positive Airway Pressure for Central Sleep Apnoea and Heart Failure (CANPAP) trial examined the effect of continuous positive airways pressure on heart failure[22]. The publication's figure 3 shows the heart transplant-free survival curve out to 60 months with an estimated transplant-free rate of 65% in the CPAP group (6 subjects still at risk) and 50% in the control with 4 subjects still at risk. The sensitivity index at 60 months is 11% for the CPAP group and 13% for the control group. The full information CI approach of Criterion 2 suggests the curve be curtailed when there is a minimum of 10 intervention-group and 7 control patients at risk. If we also require the transplant-free estimate not to decrease by more than $\Delta^* = 5\%$ in each group, the curve should be curtailed at 48 months, when $S(48) = \sim 0.64$ and $n(48) = 20$

giving a sensitivity index of 3.2% in the CPAP, and $S(48) = {\sim}0.55$ and $n(48)=19$ giving a sensitivity index 2.9% in the control group (the 48-month transplant rate is estimated at 55%).

The third example, the MA 17 trial, the overall and disease-free survival in breast cancer patients receiving an additional 5 years of letrozole after 5 years of tamoxifen or placebo [23] was investigated with a sample size over 5000 patients. The Kaplan-Meier plots were extended to 48 months, where less than 0.5% of patients were being followed up at this time. The results of these three studies are summarised in Table 2, together with how the approaches described here recommend where the Kaplan-Meier plots should be curtailed.

// Table 2: Examples of published and recommended curtailment times //

In all three cases, the Kaplan-Meier plots were extended much further than desirable from the information available, as measured by the number still at risk.

Duration of follow-up after recruitment has completed

The second data maturity issue, namely at what point during the follow-up phase of a study it would be appropriate to perform statistical analyses, is a concern in clinical studies whose outcome of interest is the survival proportion (or difference) at a pre-specified time t. As alluded to earlier the estimates at t also need to well reflect the patient population and not be unduly influenced by atypical patients. This issue becomes important particularly if (i) accrual has been prolonged; (ii) the cost of follow-up is substantial; (iii) the study result may have a major impact on clinical practice (e.g., evaluation of robotic surgery or implantation of a new medical device). Once a decision has been made to stop accrual into a study, the sample size is fixed. The only quantities which can then affect comparisons and robustness of a survival estimate at time t are the number of losses to follow-up and the number of events prior to t.

We explore these issues of data maturity by considering the *amount of actual information available*. At any time t during a study one can determine; (i) the percentage of possible information available at the current point in the study and; (ii) the increase in in statistical power (from the current power if there was no further follow-up) for comparisons if follow-up were to

continue for all subjects up until the time point of interest (e.g. 5-year survival). The variance of $S(t)$ at time t, is simply $SE(t)^2$. The ratio of the *expected $SE(t)^2$* of $S(t)$, to the current $SE(t)^2$ of $S(t)$ (as calculated say, from Greenwood's method[24]) may be used to quantify data maturity and provide a guide as to what could be expected if 100% complete follow-up at t were achieved. The *expected $SE(t)^2$* is obtained from $[S(t)(1-S(t))/N^*]$ where $S(t)$ is the current survival estimate at time t and N^* is the number of subjects who potentially can achieve full follow-up at time t (i.e., we may wish to exclude those patients who are lost to follow-up and will never be observed at t). This variance ratio will be expressed as a percentage and denoted $I(t)$. Clearly $0\% \leq I(t) \leq 100\%$ and the smaller the ratio, the less mature the data. We note that the inverse of the ratio is sometimes referred to as the variance inflation ratio.

The sample size for a study designs are typically based on 80% or 90% power and assume complete data at time t of interest. This power calculation is related to the probability of a type II error, β, and obtained from the quantiles of the standard normal distribution denoted by $z_{1-\beta}$. For 90% power, $z_{1-\beta} = 1.28$ and for 80% power $z_{1-\beta} = 0.84$. The power for a comparison of two survival proportions can be calculated as $\Pr\{Z < z(t)\}$ where Z follows a standard normal distribution and $z(t) = S(t)/SE(t)$. If we have incomplete data, $z_I(t) = z(t)\sqrt{I(t)}$. The power is fixed in advance (usually 90% or 80%), so $z(t)$ is set at either 1.28 or 0.84. For a study with no further planned accrual and in follow-up, the potential power of any comparison, if performed at the current study duration can be obtained by determining $\Pr\{Z < z_I(t)\}$ from tables of the cumulative standard normal distribution. In the examples below, we will assume that all patients entering the study are available for follow-up to the time point of interest.

To illustrate, the Z0011 trial examined the impact on 5-year survival of sentinel node biopsy (SNB) compared to auxiliary lymph node dissection (ALND) for women diagnosed with operable breast cancer [25] and aimed to enrol 1900 patients based on 90% power. The primary outcome was 5-year overall survival and the sample size was obtained by assuming a non-inferiority margin of 5% and a survival rate of 80% at 5-years, 4 years accrual and 5-years follow-up. The study closed early due to a low event rate, enrolling only 856 patients, after 5.6 years accrual and 5.1 years

follow-up and having a pooled 5-year survival rate of 92.2%. Based on this duration, a sample size of 856, a pooled 5-year survival rate of 92% and 5% non-inferiority margin, the power of the comparison has reduced from 90% to 67%. If we consider just the 420 patients allocated to ALND, the estimated 5-year rate was 91.8% (95% CI, 89.1%-94.5%) with 313 patients still at risk at this time and known not to have died. The *expected SE(t)^2* is (0.918*0.082)/420 and the *SE(t)^2* of the estimate at 5-years is 0.01378^2 giving $I(5)$ = 94% with a similar result in the SNB group. Of the 52 deaths in this group, 32 deaths would be expected to have occurred prior to 5-years, giving the rate of completeness of follow-up as 345/420= 82%. For 67% power, $z_{1-\beta}$ = 0.44 and $z_1(t)$ = 0.44$\sqrt{0.94}$ so, performing the comparisons with the current follow-up would give a power of Pr{Z <0.426) or 66.5%, a minor decrease from 67%.

While we have used just the ALND group to illustrate these ideas, in practice, all these quantities should only be obtained from pooled data (blinded to treatment allocation) to ensure no bias is introduced into the decision as to whether the data should be analysed at the current study duration. It is crucial that investigators and statisticians remain blinded to treatment allocation prior to these decisions being made. Decisions regarding study accrual/follow-up should be made prior to study commencement and decisions to stop a study prior to the accrual being met should be made by an independent data safety and monitoring committee who would assess the totality of the evidence before making their recommendations. Such committees would also review pooled data before requesting unblinded information.

We can also apply these ideas to the I-ELCAP study. The median follow-up for the observation group is 46 months (based on the reverse Kaplan-Meier method[26]) and, if the outcome of interest is the percentage of patients surviving at 72-months, the survival estimate is 81% with a variance of 0.0215^2. The full information variance at 72-months is (0.81*0.19)/484 (assuming no losses to follow-up), $I(72)$ is 69% and $I(120)$ is 41%. We note that 51 out of 410 patients (12.2%) not having an event have been followed to 72 months and 2 out of 409 (0.5%) have been followed to 120 months. In this example where the prime interest in the precision of the $S(t)$, the number at risk at t together with $I(t)$ will help the interpretation of the maturity of follow-up at these times.

The change in the change in study power based on different levels of $I(t)$ for study designs based power of 80% 85% and 90% is shown in figure 3.

// Figure 3 Expected power for different levels of $I(t)$ for studies initially designed to have 90%, 85% and 80% power with full follow-up.//

For studies with sample sizes based on power of 90%, $I(t)$ as low as 45% would still provide adequate power (80%) for planned comparisons.

We can use the number at risk as defined by criterion 1 or 2 together with the potential power of any planned comparisons (if appropriate), to provide a guide as to whether there is sufficient data maturity to stop follow-up. If at time t the actual number at risk satisfies the criteria (1 and/or 2) of choice, then this fact together with $I(t)$ provides a guide as to the impact of stopping follow-up at this time. For data from population registries where high precision of the estimates is required, $n(t)$ should satisfy criterion 1 together with a high value of $I(t)$ to ensure the robustness of the published estimates.

Discussion

Two related problems associated with data maturity in clinical studies having time-to-event outcomes have been explored; (i) difficulty of drawing sensible conclusions from published displays where the survival estimate is sensitive to an extra event making clinical decisions difficult; and (ii) how much follow-up on completion of study recruitment is required prior to reporting of the study results. While much of the focus has been on the impact of the number at risk and the precision of the survival estimate at time t the issue of how well these patients represent the disease population should also be considered. Survival curves extended to the last known event time (so called "all the way"[7], a view held by many investigators) may well be misleading, cloud interpretation and compromise the generalisability of the study.

We defined the sensitivity index $\Delta(t)$ of the Kaplan Meier estimate $S(t)$ at a t, as the decrease of the %survival estimate had one extra event occurred immediately post t. We propose two criteria

to assess $\Delta(t)$. The first sets a maximum acceptable threshold Δ^* for the sensitivity index in the context of the investigation. The second criterion restricts $\Delta(t)$ to be no larger than the width of the full information one-sided 95% CI for the %survival at t. These approaches assist investigators in deciding when to curtail the Kaplan-Meier plot and so avoid potential misrepresentation of the survival estimates. The 95% confidence interval at t is commonly obtained from Greenwoods formula but this approach has limitations [9, 12]. When compared to the recommendation of Pocock[7], the proposed strategies are less conservative.

More recently Fay et al.[11] have proposed a method which adjusts the width of the confidence interval to account for changes in the number at risk based on the product of beta random variables. While the method is methodologically complex, with calculations currently only available in the statistical package R, it nevertheless illustrates the increase in uncertainty with increased censoring. However, when there are multiple curves being presented (as in the ADAPT study[27]), the benefit of the visual representation can be masked with the multiple confidence bands. Fay et al's lower one-sided 95% CI is shown as the shaded area in figure 1; the lower limit at 120 months being 0.21 and, at 84 months, 0.70. These values are larger than 0.76, the value given at both times by the Greenwood method. While these intervals demonstrate the uncertainty in the estimates they do not provide guidelines as to when survival plots should be curtailed. Our approach provides a framework for interpretation of survival estimates at specified time points based on both the uncertainty and the number at risk at these points.

This approach has applicability to a wide range of practical problems, and does not rely on assumptions of specific censoring patterns of the trial or the extent of censoring present in the data. These guidelines have applications in deciding on the time point at which the tail of the survival curve should be modelled in studies of cost-effectiveness and quality-adjusted survival. With registry or population information, the sensitivity index of the Kaplan-Meier plot to individual events should be low. Such precision is demanded by health decision makers when establishing policy guidelines. When interventions are evaluated for effectiveness over long periods (radiotherapy, surgical procedures, and implanted devices), it is essential the disease or

device failure-free rates at 3, 5, or 10 years are not sensitive to an additional event. In these cases, the decrease $\Delta(t)$ in the % survival if one extra event were to occur should be small (1% or, at most 2.5%) to provide reasonable precision boundaries. Different levels of precision may be desired/warranted depending on the problem being investigated.

The second problem related to data maturity, examines the completeness of follow-up once accrual to a study has finished. This is based on the amount of actual information available relative to having complete follow-up at time t. We use this quantity to calculate the potential statistical power of any comparisons providing a guide as to whether an individual study has sufficient data maturity at the current duration of follow-up.

Bias induced by differential censoring has been studied by Beltangady *et al* for the log-rank test[28] and by Persson *et al* [29] for the proportional-hazards model. This supports the need for guidelines for (a) when survival estimates should be quoted and when the corresponding comparisons of these estimates at key time-points can sensibly be performed, and (b) how much follow-up should be planned. These issues are crucial when reporting results of observational studies, especially from population disease registries or clinical databases which provide rates of clinical benefit or toxicities etc., to inform future studies or help formulate public health policy.

Conclusion

We provide a simple and effective framework to gauge the maturity of data at different follow-up times to inform researchers as to which follow-up time provides sufficient information to allow sensible interpretation of survival estimates. Two approaches for when to curtail the Kaplan-Meier plot are proposed, both related to the decrease in the % survival estimate if one further event were to occur. The minimum required number at risk at any time point based on the full information CI has a straightforward interpretation and can be easily calculated from the output provided by common statistical packages. We provide an approach to ascertain the actual information based on the current number of events and current follow-up times in a clinical study. This allows investigators to determine the statistical power certain comparisons may have if

analyses were performed at this time. For non-comparative studies the percentage of actual information will help inform whether follow-up is sufficient or needs to be prolonged.

Funding

Acknowledgement

References

1. Bland M, Altman D. Survival probabilities (the Kaplan-Meier method). *Br Med J* 1998; **317**: 1572.

2. Peto J. The calculation and interpretation of survival curves. In: Buyse M, Staquet M, Sylvester R, editors. *Cancer Clinical Trials Methods and Practice*. Oxford: Oxford University Press; 1984. p. 361–80.

3. Mallick S, Benson R, Rath G. Patterns of care and survival outcomes in patients with pineal parenchymal tumor of intermediate differentiation: An individual patient data analysis. *Radiother Oncol* 2016; **121**: 204–8.

4. Matthay K, Villablanca J, Seeger R, 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.

5. Herbert C, Liu M, Tyldesley S, et al. Biochemical Control With Radiotherapy Improves Overall Survival in Intermediate and High-Risk Prostate Cancer Patients Who Have an Estimated 10-Year Overall Survival of >90%. *Int J Radiat Oncol Biol Phys* 2012; **83**: 22-7.

6. Carter R, Huang P. Cautionary Note Regarding the Use of CIs Obtained From Kaplan-Meier Survival Curves. *J Clin Oncol* 2009; **27**: 174-5.

7. Pocock S, Clayton T, Altman D. Survival plots of time-to-event outcomes in clinical trials: Good practice and pitfalls. *Lancet* 2002; **359**: 1686–9.

8. Clark T, Altman D, De Stavola B. Quantification of the completeness of follow-up. *The Lancet* 2002; **359**: 1309–10.

9. Borkowf C. A simple hybrid variance estimator for the Kaplan–Meier survival function. *Stat Med* 2005; **24**.

10. Fay M, Brittain E. Finite sample pointwise confidence intervals for a survival distribution with right-censored data. *Staistics in Medicine* 2016; **35**: 2726–40.

11. Fay M, Brittain E, Proschan M. Pointwise confidence intervals for a survival distribution with small samples or heavy censoring. *Biostatistics* 2013; **14**: 723–36.

12. Miettinen O. Survival analysis: up from Kaplan–Meier–Greenwood. *Eur J Epidemiol* 2008; **23**: 585–92.

13. Murray S. Using Weighted Kaplan-Meier Statistics in Nonparametric Comparisons of Paired Censored Survival Outcomes. *Biometrics* 2001; **57**: 361-8.

14. Rossa A, Zieliński R. A simple improvement of the Kaplan-Meier estimator. *Communications in Statistics - Theory and Methods* 2002; **31**: 147-58.

15. Gray B, Van Hazel G, Hope M, et al. Randomised trial of SIR-Spheres((R)) plus chemotherapy vs. chemotherapy alone for treating patients with liver metastases from primary large bowel cancer. *Ann Oncol* 2001; **12**: 1711-20.

16. Wasan H, Gibbs P, Sharma N, et al. First-line selective internal radiotherapy plus chemotherapy versus chemotherapy alone in patients with liver metastases from colorectal cancer (FOXFIRE, SIRFLOX, and FOXFIRE-Global): a combined analysis of three multicentre, randomised, phase 3 trials. *Lancet Oncolgy* 2017; **Published Online August 3, 2017 http://dx.doi.org/10.1016/S1470-2045(17)30457-6**.

17. Mariette C, Dahan L, Mornex F, et al. Surgery Alone Versus Chemoradiotherapy Followed by Surgery for Stage I and II Esophageal Cancer: Final Analysis of Randomized Controlled Phase III Trial FFCD 9901. *J Clin Oncol* 2014; **32**: 2416-22.

18. Leong T, Smithers M, Michael M, et al. TOPGEAR: a randomised phase III trial of perioperative ECF chemotherapy versus preoperative chemoradiation plus perioperative ECF chemotherapy for resectable gastric cancer (an international, intergroup trial of the AGITG/TROG/EORTC/NCIC CTG). *BMC Cancer* 2015; **15**.

19. Janda M, Gebski V, Davies S, et al. Effect of Total Laparoscopic Hysterectomy vs Total Abdominal Hysterectomy on Disease-Free SurvivalAmong Women With Stage I Endometrial Cancer A Randomized Clinical Trial. *Journal of American Medical Association* 2017; **317**: 1224-33.

20. Investigators TIELCAP. Survival of Patients with Stage I Lung Cancer Detected on CT Screening. *N Engl J Med* 2006; **355**: 1763-71.

21. Vilgrain V, Pereira H, Assenat E, et al. Efficacy and safety of selective internal radiotherapy with yttrium-90 resin microspheres compared with sorafenib in locally advanced and inoperable hepatocellular carcinoma (SARAH): an open-label randomised controlled phase 3 trial. *Lancet Oncol* 2017; **Published Online October 26, 2017**: http://dx.doi.org/10.1016/S470-2045(17)30683-6.

22. Bradley T, Logan A, Kimoff G, et al. Continuous Positive Airway Pressure for Central Sleep Apnea and Heart Failure. *N Engl J Med* 2005; **353.**: 2025-33.

23. Goss P, Ingle J, Martino S, et al. A Randomized Trial of Letrozole in Postmenopausal Women after Five Years of Tamoxifen Therapy for Early-Stage Breast Cancer. *N Engl J Med* 2003; **349**: 1793-802.

24. Greenwood M. The natural duration of cancer. Reports on Public Health and Medical Subjects. *Her Majesty's Stationery Office, London* 1926; **33**: 1-26.

25. Giuliano A, Hunt K, Ballman K, et al. Axillary Dissection vs No Axillary Dissection in Women With Invasive Breast Cancer and Sentinel Node Metastasis A Randomized Clinical Trial. *Journal of American Medical Association* 2011; **305**: 569-75.

26. Korn E. Censoring distributions as a measure of follow-up in survival analysis *Stat Med* 1986; **5**: 255-60.

27. ADAPT Research Group. Cardiovascular and Cerebrovascular Events in the Randomized, Controlled Alzheimer's Disease Anti-Inflammatory Prevention Trial (ADAPT). *PLoS Clin Trials* 2006; **1(7): e33**.

28. Beltangady M, Frankowski R. Effect of unequal censoring on the size and power of the logrank and Wilcoxon types of tests for survival data. *Stat Med* 1989; **8**: 937-45.

29. Persson I, Khamis H. Bias Of The Cox Model Hazard Ratio. *Journal of Modern Applied Statistical Methods* 2005; **4**.

Key Messages

- The information provided by the Kaplan-Meier (KM) curve at a particular time point is dependent on the number of subjects at risk at this time point. If there are only a few patients at risk, then one single extra event will make a substantial impact on the distance by which the KM curve decreases.

- Data maturity is explored with respect to two problems: (i) how far in time a KM curve should be extended and (ii) at what point in a clinical study, prior to achieving complete follow-up, can it still be appropriate to report results.

- For (i), a minimum desirable number at risk is readily obtained if the drop in the survival estimate at time t should one extra event occur is required to be less than a pre-determined threshold. To ensure this drop remains below the threshold, large survival estimates require more patients remaining at risk than smaller ones.

- An alternative approach to (i) is to require that the decrease in the survival estimate at time t should one extra event occur does not exceed the width of the one-sided 95% CI for survival based on 'full information' if all the data were available up to the time of interest.

- For (ii), when follow-up is incomplete at some time point, the variance of the KM estimate is larger than if there had been complete follow-up. The variance ratio represents the proportion of actual information available at the time point. It can be used in comparative studies to determine the potential statistical power if study results are reported for that time point without further follow-up.

- The amount of current information available and the percentage of complete follow-up can be used to both plan the follow-up duration of the current study to ensure that a study attains an acceptable level of actual information or sufficient expected power prior to analysis of the results.

Figure legend

Figure 1

Kaplan-Meier plot for International Early Lung Cancer Investigators I-ELCAP Study.

— · — The impact on the estimate if one extra event at 120 months were to occur

······► Curtail the plot the 20% (60 months) point (Pocock rule)

— — ► Curtail the plot the 10% (72 months) point (Pocock rule)

———► Curtailment point for the full follow-up one-sided 95% CI criteria (84 months).

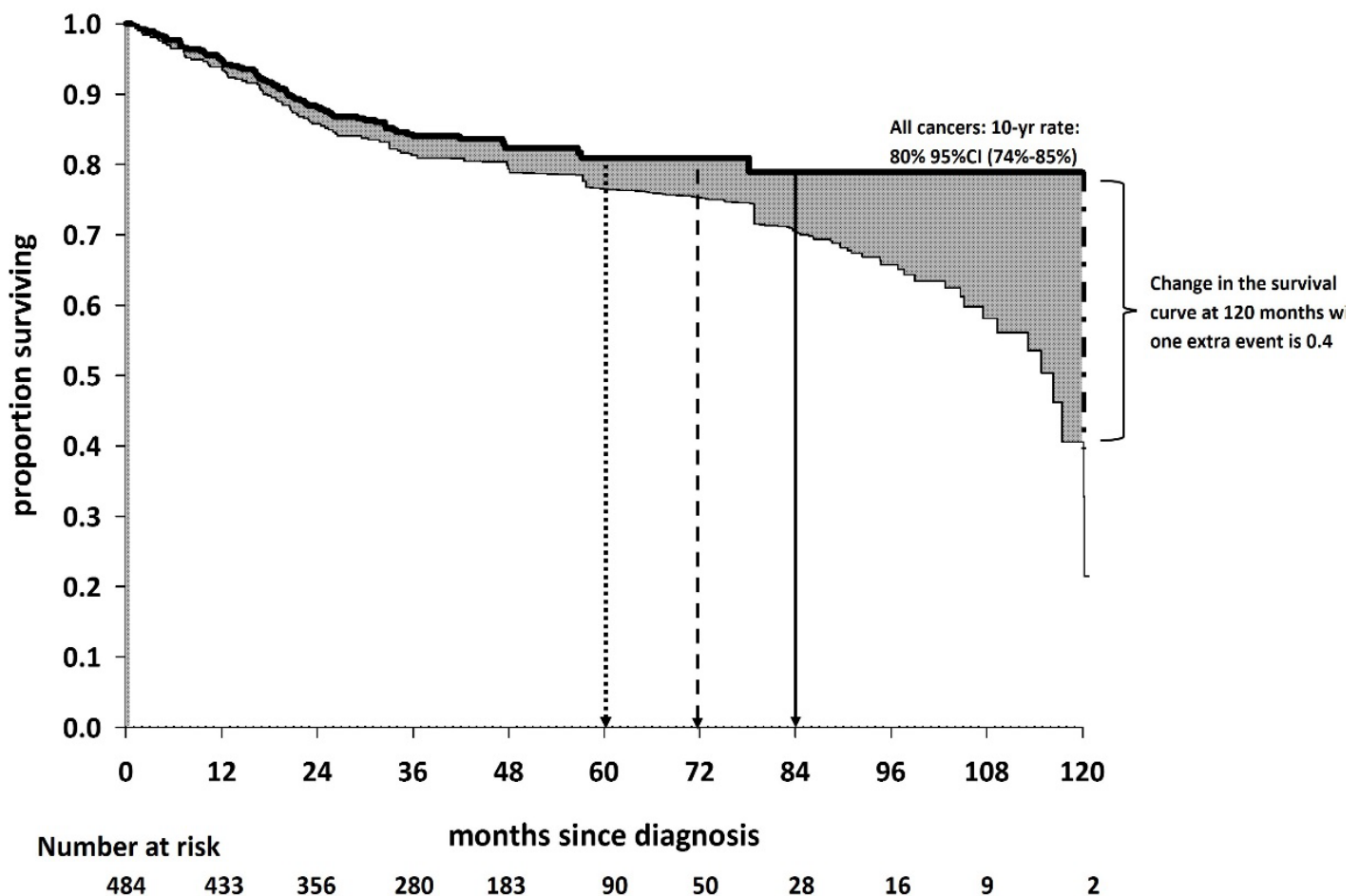Shaded area: Fay and Brittain 95% one-sided pointwise confidence bands

Figure 2

Minimum number of patients at risk for different sample sizes for values of the Kaplan-Meier curve, $S(t)$ from 0.1 to 0.95
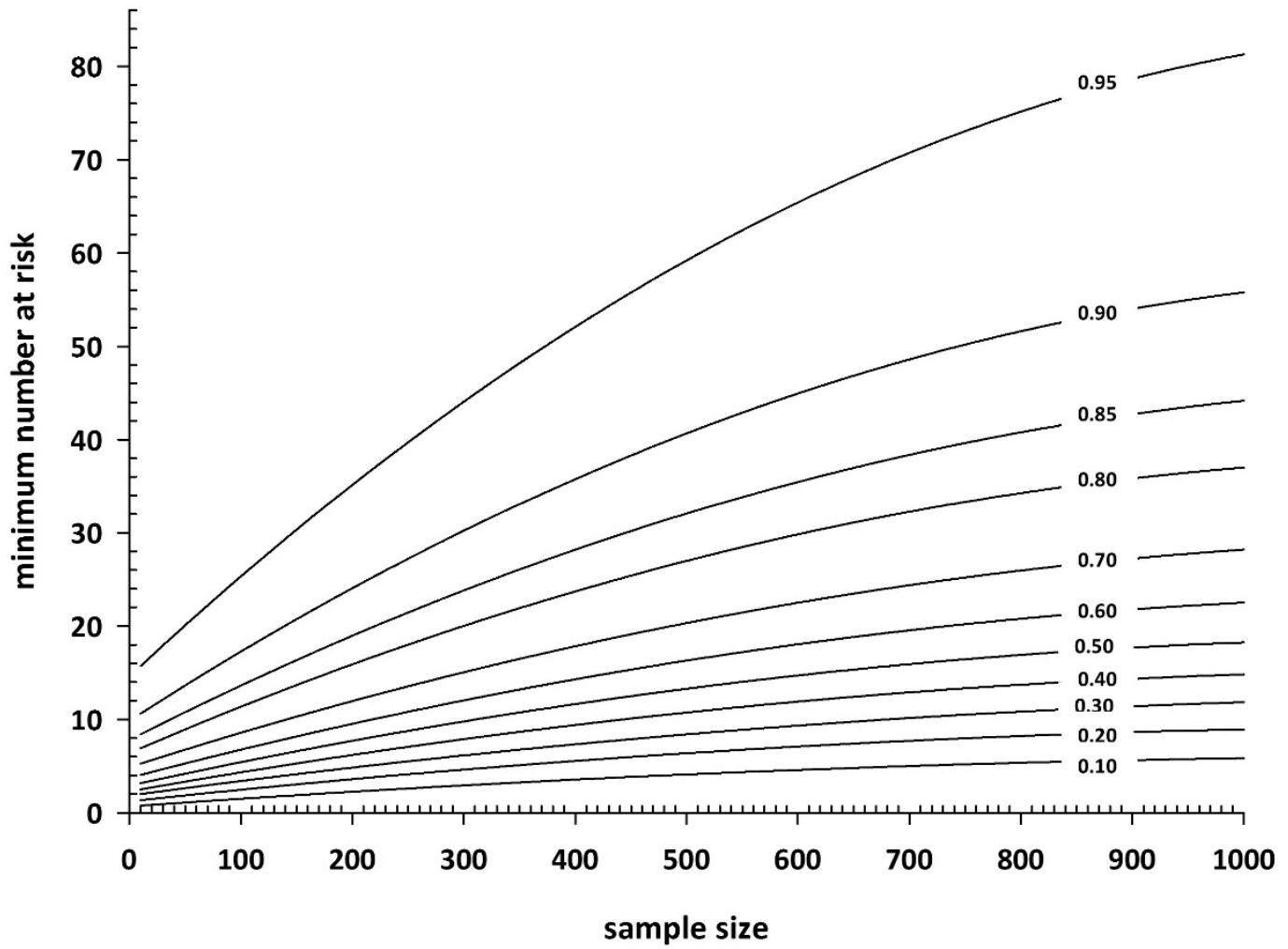
Figure 3

Levels of statistical power for percentages of actual information available in study designs having 80%, 85% and 90% power with complete information
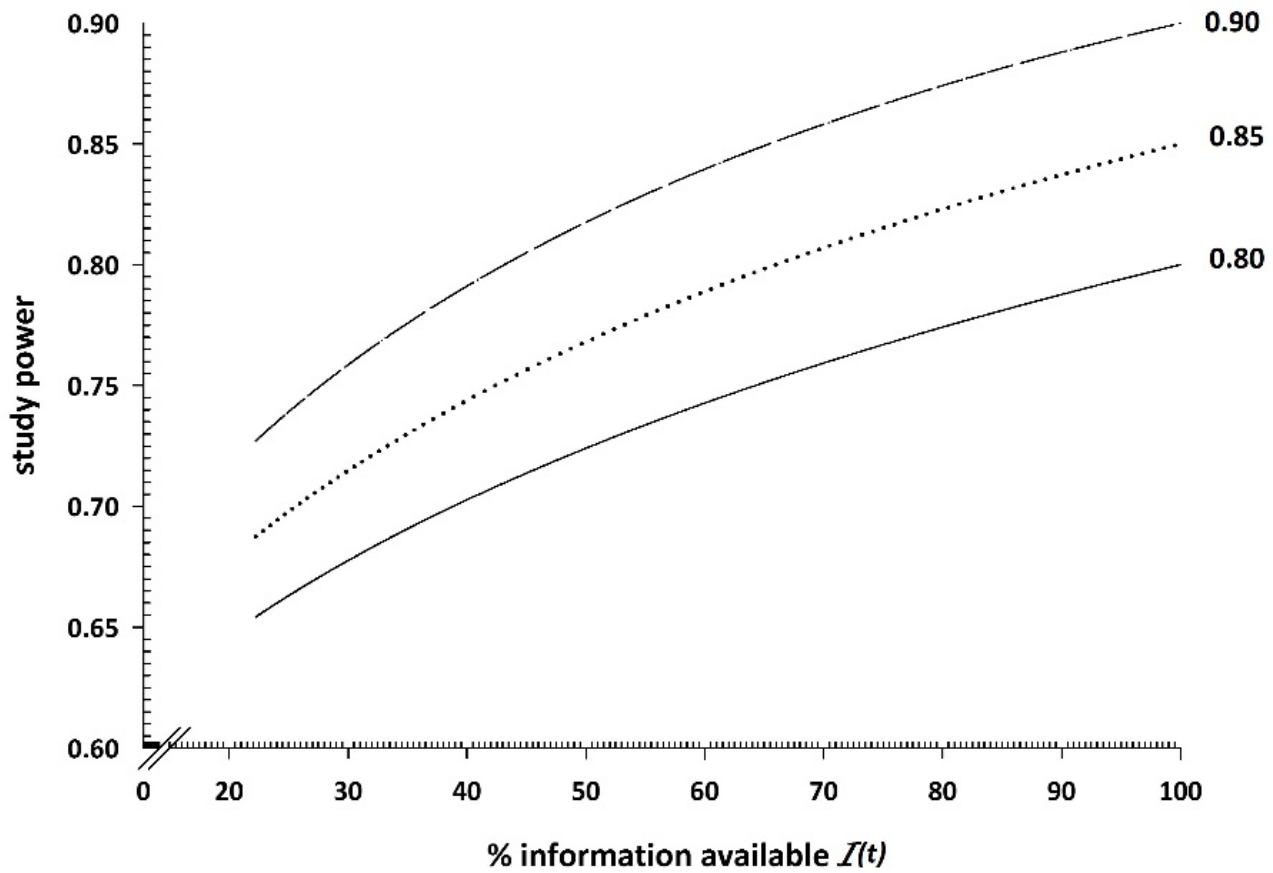
Table 1: Number at risk of an event in the I-ELCAP lung study

Time	Survival estimates	Actual number at risk (n_r)	Minimum n satisfying criretrion 2	decrease in the %survival estimate for one extra event[‡]
0	1.00	484		0.21
6	0.98	456	87	0.21
12	0.95	434	58	0.22
18	0.92	390	45	0.24
24	0.88	357	37	0.25
30	0.86	322	34	0.27
36	0.84	281	31	0.30
42	0.84	236	31	0.35
48	0.82	184	29	0.45
54	0.82	133	29	0.62
62	0.81	91	28	0.89
66	0.81	67	28	1.21
72	0.81	51	28	1.59
78	0.81	41	28	1.97
84	0.79	29	26[*]	2.72[†]
90	0.79	21	26	3.76
96	0.79	16	26	4.93
102	0.79	11	26	7.17
108	0.79	9	26	8.77
114	0.79	7	26	11.27
120	0.79	2	26	39.50

[*]Number at risk according to criterion 2 required for extending the survival curve based on full information. If this minimum is larger than the actual number at risk, then not extending the curve past this time point should be considered

$$^{†}26 = \frac{1}{1.645}\sqrt{\frac{484*(0.79)}{(1-0.79)}}$$

[‡]100*(0.79/29)% is the decrease in the % survival;estimate if one extra event were to occur at 84 months.

Table 2: Examples of published and recommended curtailment times.

Study	I-ELCAP	CANPAP		MA 17	
Outcome	Overall survival	Transplant-free survival		Disease-free survival	
		CPAP	Control	Control	Letrozole
Published study results					
Total sample size	484	128	130	2582	2575
Curtailment time	120 months	60 months		48 months	
At published curtailment time					
$100S(t)$	80%	65%[*]	50%[*]	83%[*]	93%[*]
Number at risk	2	6	4	11	9
Sensitivity index[‡]	40%	11%	13%	8%	10%
Recommended curtailment time					
Minimum number at risk based on Criterion 2	27	10	7	71	149
Corresponding curtailment time (months)	84	48		42[†*]	
Sensitivity index	2.72%	3.2%	2.9%	0.66%	0.64%
Pocock's 10% (20%) rule					
Number at risk from the 10% (20%) rule	48 (97)	13 (26)		258 (516)	
Curtailmant time (months) at 10% (20%)	72 (60)	54(42)		38(32)	

† assumes uniform attrition between 40 and 50 months of 17 patients per month in each group; $S(42) = 0.96$ for Letrozole and $S(42)= 0.85$ for control

[*] Estimated from the published figure.

[‡] at the published curtailment time