Interim Analyses of Survival Data in Cancer Clinical Trials

Eva Skovlund

From the Norwegian Cancer Society and Section of Medical Statistics, University of Oslo, Oslo, Norway

Correspondence to: Dr Eva Skovlund, The Norwegian Radium Hospital, Montebello, N-0310 Oslo, Norway. Tel: +47 22 93 47 56. Fax: +47 22 50 91 99. E-mail: eva.skovlund@klinmed.uio.no

Acta Oncologica Vol. 37, No. 7/8, pp. 645–650, 1998

It is common practice to undertake interim analyses of accumulating survival data in a cancer clinical trial while patient entry and/or follow-up is still in progress. The purpose of making an interim analysis is to stop the trial if a convincing treatment difference can be demonstrated. The main problem with repeated significance tests is that the more often one analyses accumulating data, the greater the probability of eventually obtaining a significant result, $p < 0.05$ say, even when there is in reality no treatment difference. To allow for repeated testing one must therefore choose a more stringent nominal significance level as a criterion for stopping the trial. Repeated significance tests are normally applied at equally spaced intervals, and the maximum number of tests is decided in advance. This paper presents and compares the properties of two simple and commonly applied strategies for undertaking interim analyses of accumulating survival data.

Received 4 May 1998
Accepted 28 September 1998

A major issue in clinical cancer research is the comparison of patient survival or time to recurrence on different treatments. Patients are usually included in a trial over a long period, and thereafter they may be followed for several years. The number of patients to be recruited in a clinical trial as well as the length of follow-up is usually decided before a trial is initiated.

Most medical studies are analysed by performing significance tests. A p-value is calculated and compared with a specified significance level, usually 0.05. The significance level of a test is the probability of obtaining a false positive result; i.e. concluding that one treatment is better than another when they are in fact equal. This is commonly referred to as a type I error. If a p-value is less than 0.05, it is usually concluded that treatment effects differ. A large p-value is, on the other hand, regarded as evidence that treatment effects are similar. The latter conclusion may not always be valid. With few patients in a trial the probability of not detecting a treatment difference even when treatment effects are in reality different may be large. It is therefore important to estimate the sample size required in order to detect clinically important differences in treatment effect before starting a trial.

During the period of inclusion and follow-up of patients, data on time to recurrence and overall survival accumulate gradually. For ethical reasons it may be of interest to stop a trial once a significant difference between treatments can be demonstrated, and natural curiosity to possible differences in treatment effects may make researchers want to analyse data before the planned follow-up has been completed. These early analyses are called interim analyses and their main objective is to look for treatment differences that are sufficiently convincing to stop the trial. Interim analyses are usually made by performing repeated significance tests at equally spaced intervals. Unfortunately, it does not seem to be generally appreciated that unplanned interim analyses could substantially increase the risk of a false positive result.

When interim analyses are carried out, the significance level for each analysis must be adjusted so that the desired overall probability of a false positive result is maintained (usually at 0.05). Without some kind of adjustment of the significance level for each repeated test, the overall significance level will be higher than planned, and the risk of drawing a false positive conclusion may become unacceptably high (1). The more often accumulating data are analysed, the greater the risk of eventually rejecting the null hypothesis of no treatment difference, even when the null hypothesis is in reality true. Choosing a more stringent nominal significance level for each repeated test can solve the problem. Such methods are often referred to as group sequential methods.

Several different methods for interim analyses have been proposed. The purpose of this paper is to present and

 ISSN 0284-186X

compare two relatively simple methods for interim analyses, which are commonly used in clinical trials (2, 3). The methods are compared both by simulating survival data from exponential distributions and by reanalysing a real data set. The effect of making unplanned interim analyses without adjusting the significance level for each repeated test is also illustrated. Recommendations regarding the choice of method are presented.

STOPPING RULES FOR INTERIM ANALYSES

Planned interim analyses should be simple, usually involving only the primary endpoint. The maximum number of analyses should be determined in advance. Interim analyses of survival data are usually carried out at regular intervals, most often equally spaced according to number of events (e.g. deaths or recurrences). At each interim analysis a significance test is performed. If the p-value is smaller than some nominal significance level, which is specified in advance, the trial is stopped; otherwise the trial continues until the next planned interim analysis. For practical reasons it may be preferable to analyse data at certain calendar times. It has been shown that group sequential methods are robust to variation in the timing of the analyses (4).

Common strategies for interim analyses are different adjustments of the nominal significance level. The nominal level is chosen such that the desired overall significance level (e.g. 0.05) is maintained. Pocock (2) suggests a strategy based on a fixed nominal level. O'Brien & Fleming (3) propose an alternative design with different nominal level for each interim analysis. Values of nominal levels when a maximum of 1 to 5 analyses is planned are presented in Table 1. Large gains are expected by planning for a maximum of two analyses instead of one. Pocock (5) has shown, however, that relatively little is gained by conducting more than three to five analyses.

SIMULATION METHODS

The properties of different strategies for interim analysis are compared by stochastic simulation. The simulation programs were written in SIMULA (6) and executed on a SUN computer at the University of Oslo.

The increase in overall significance level by carrying out unplanned interim analyses was estimated by sampling survival data from exponential distributions. Each simulated patient was randomly allocated to one of the two groups with equal probability ($p_{treat} = p_{control}$). Under the null hypothesis of no treatment difference responses were sampled from identical distributions. For estimations of power, responses were sampled from two exponential distributions with unequal hazard, i.e. unequal 'risks of death'. A set of n observations was thus generated by randomly drawing approximately n/2 patients from each of the two distributions. The simulated observations were sorted by time of inclusion, and all patients were followed until a defined number of events had occurred. Individuals still alive were treated as censored observations. Simulated survival times were compared between the two groups by the log-rank test. If the stopping criterion for a certain analysis strategy was reached, the trial was stopped for that particular method, otherwise the trial continued until the next planned interim analysis. Again, the result of each analysis strategy was registered and the trial was stopped or continued depending on the result. This was repeated until the planned number of events had occurred and the planned number of analyses had been performed. Each such sequence was repeated 10000 times. By approximation to the normal distribution it can be shown that with 10000 simulations, estimates of overall significance level outside the interval (0.0464, 0.0536) are significantly different from 0.05.

The interim analyses were carried out at equally spaced intervals in terms of number of events. The simulation results for each strategy are based on identical data. Differences between strategies are thus not due to random variation.

TRUE OVERALL SIGNIFICANCE LEVEL

If repeated significance tests are performed without applying a more stringent nominal level at each interim analysis, the probability of a false positive conclusion will be much higher than desired. The overall significance level obtained when unplanned interim analyses are performed is plotted in Fig. 1. The estimated level increases with increasing number of analyses. With a maximum of five analyses the

Table 1

Nominal significance levels for interim analyses with overall significance level 0.05

Max. no. of analyses	Analysis no.	Nominal significance level	
		Pocock	O'Brien & Fleming
1	1	0.05	0.05
2	1	0.029	0.0051
	2	0.029	0.0475
3	1	0.022	0.0006
	2	0.022	0.0151
	3	0.022	0.0472
4	1	0.018	0.00004
	2	0.018	0.0039
	3	0.018	0.0184
	4	0.018	0.0411
5	1	0.016	0.000005
	2	0.016	0.0013
	3	0.016	0.0085
	4	0.016	0.0228
	5	0.016	0.0417

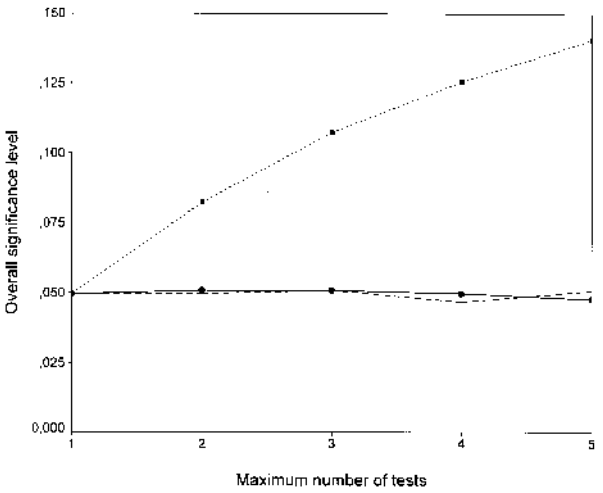

Fig. 1. Estimated overall significance level. Filled squares and dotted line represent unadjusted tests (each at the 0.05 level), filled circles and solid line Pocock's method, and open circles and dashed line O'Brien & Fleming's method. The desired overall significance level is 0.05. Each estimate is based on n = 10000 simulations.

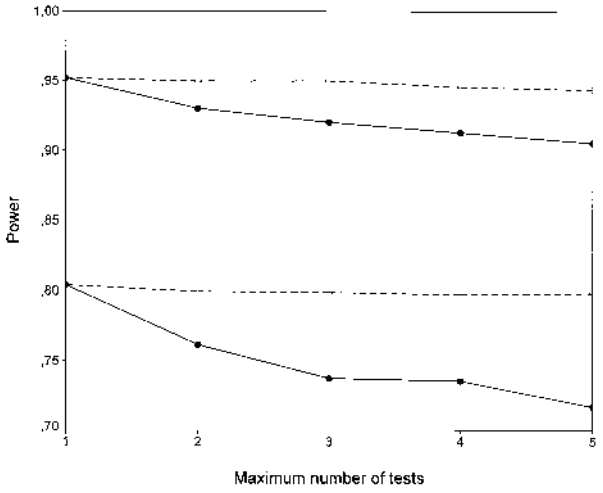

Fig. 2. Estimated power. Filled circles and solid line represent Pocock's method and open circles and dashed line O'Brien & Fleming's method. The desired power is 0.80 and 0.95, respectively. Each estimate is based on n = 10000 simulations.

true level is about 0.14 compared with the desired level of 0.05. The two methods adjusting the nominal level for each interim analysis obtain an overall level of 0.05, as desired.

SAMPLE SIZE AND POWER

When a clinical trial is planned, the number of patients necessary to draw a valid conclusion must be estimated. Table 2 illustrates the number of patients that need to be included in order to detect some clinically relevant differences in treatment effect. It is assumed that the 5-year survival in the control group is 40%. Assuming constant hazard (i.e. exponential survival), this corresponds to a median survival of 45 months. Patient accrual is set to 4 years and the longest follow-up possible to 10 years. The table shows the total number of patients necessary to obtain a power of 0.80 and 0.90, respectively. The patient numbers have been calculated assuming that a trial with fixed sample size is planned. In order to uncover small treatment differences, large patient numbers are needed.

The two methods are both capable of maintaining the desired overall significance level, but they differ concerning power to detect treatment differences. This is illustrated in Fig. 2. The O'Brien and Fleming method demonstrates a negligible loss of power with an increasing maximum number of tests. With Pocock's fixed nominal level method, however, the power decreases markedly with an increasing maximum number of tests. With a maximum of five analyses the true power is reduced from 0.80 to 0.72 if the patient number necessary for a fixed sample trial is allocated. The loss of power is smaller when the planned power is higher. If the planned power is 0.95, the actual power with Pocock's method will be 0.91 when a maximum of five analyses is allowed.

Table 2

Number of patients required to detect various survival differences

5-year survival in treatment group[1]	Median survival in treatment group[2]	Hazard ratio[3] control/treatment	Total number of patients	
			Power 0.80	Power 0.90
0.70	117	2.57	64	86
0.65	97	2.13	92	124
0.60	81	1.80	142	190
0.55	70	1.53	256	342
0.50	60	1.32	576	770
0.45	52	1.15	2 164	2 896

[1] 5-year survival in control group 0.40.
[2] Median survival in control group 45 months.
[3] Often referred to as relative risk.

In order to maintain the desired power, the maximum number of patients must be increased when Pocock's method is applied. The approximate increase required to maintain the power is shown in Table 3 (2).

UNPLANNED INTERIM ANALYSES

Many cancer clinical trials are planned with a fixed sample size, but may turn out to require one or more unplanned interim analyses during the course of the trial. It is then possible to redesign the trial before any analysis has been undertaken. The O'Brien and Fleming method is especially amenable since the first analysis would have a very stringent nominal level. It also has the advantage of maintaining the desired power. Another possibility would be to undertake repeated analyses at the 0.001 level. The final analysis could then be conducted at nearly the 0.05 level. An overall level of 0.05 would still be maintained unless many interim analyses were conducted (7).

AN EXAMPLE

A real data set has been used to compare the properties of the different methods. The data set consists of 350 patients with estrogen receptor positive, operable, axillary node positive breast cancer (8). Patients were randomized to receive tamoxifen (TAM) 20 mg/day for 2 years (180 patients) or no adjuvant hormone therapy (170 patients). Patients were included over a period of 5 years. Data on recurrence-free survival are used for illustration. The trial was performed with a fixed sample size and no interim analyses were in reality conducted.

What might have been the result of performing interim analyses during this trial? A possible design could have been to plan for a maximum of three analyses. The timing of the analyses might be approximately 4, 7 and 10 years after the first patient was entered. We assume that a maximum of 350 patients is entered. Estimated survival after 4, 7 and 10 years, respectively, is illustrated in Fig. 3.

After 4 years, 266 patients had been enrolled; 140 on TAM and 126 in the control group. Of these patients, 22 and 34 had relapsed, respectively, and the log-rank test gave $p = 0.0114$. This p-value is lower than Pocock's stopping rule (nominal level $\alpha' = 0.022$), so the trial would have been terminated had this method been used. Since the p-value is larger than the O'Brien and Fleming stopping rule (nominal level $\alpha' = 0.0006$), the trial would have continued until the next planned analysis at 7 years if this method had been chosen.

Seven years after start of the study 53 of the 180 patients receiving TAM had relapsed, compared with 88 of the 170 patients in the control group. The log-rank test would have given $p = 0.0001$, which is lower than the nominal level ($\alpha' = 0.0151$), and the trial would have been terminated. The results of employing the two group sequential methods and the fixed sample trial actually performed are compared in Table 4.

Table 3

Required increase in maximum patient number with Pocock's method

Maximum number of analyses	Required increase (%)
2	10
3	15
4	18
5	20

DISCLOSURE OF INTERIM RESULTS

Who should review the data and decide when to recommend termination of a clinical trial? Most cancer cooperative groups have established data safety monitoring committees to monitor interim results and recommend to the group when a study should be altered, terminated, or reported. The main reason for establishing these committees is patient safety.

In the past, interim analyses and survival curves were often available to all participating physicians. This practice may have a considerable influence on the future progress of a trial. If an early interim analysis shows little difference, physicians may lose interest and enter few, if any, new patients in the trial. An even more serious situation would arise if there were interesting but non-significant differences. Some physicians may stop randomizing patients, as they believe there is a true difference. Others may include only a subgroup of eligible patients. If the trial is not double blind, there is also a risk that patients on one of the treatments will be withdrawn, usually because the physician believes they are receiving inferior treatment. This may introduce selection bias.

Many trials have terminated prematurely because patient entry has stopped as a result of the disclosure of interim data. An incomplete picture of treatment effects will be the result, and it is strongly recommended that no presentation or publication of interim results should take place while patients are still being randomized (9). Premature publication may have the effect of the whole medical community being prejudiced towards a particular conclusion before the full results are known.

DISCUSSION

When results are properly interpreted, interim analyses will reduce the average time required to conduct a trial without increasing the probability of obtaining a false positive result. If no adjustment of the significance level is undertaken, there will be a serious increase in the probability of reaching a false positive conclusion. This misinterpretation of the strength of evidence when repeated analyses are conducted over time may partly explain why many published studies that were terminated early with apparently significant results have not been reproduced (10).

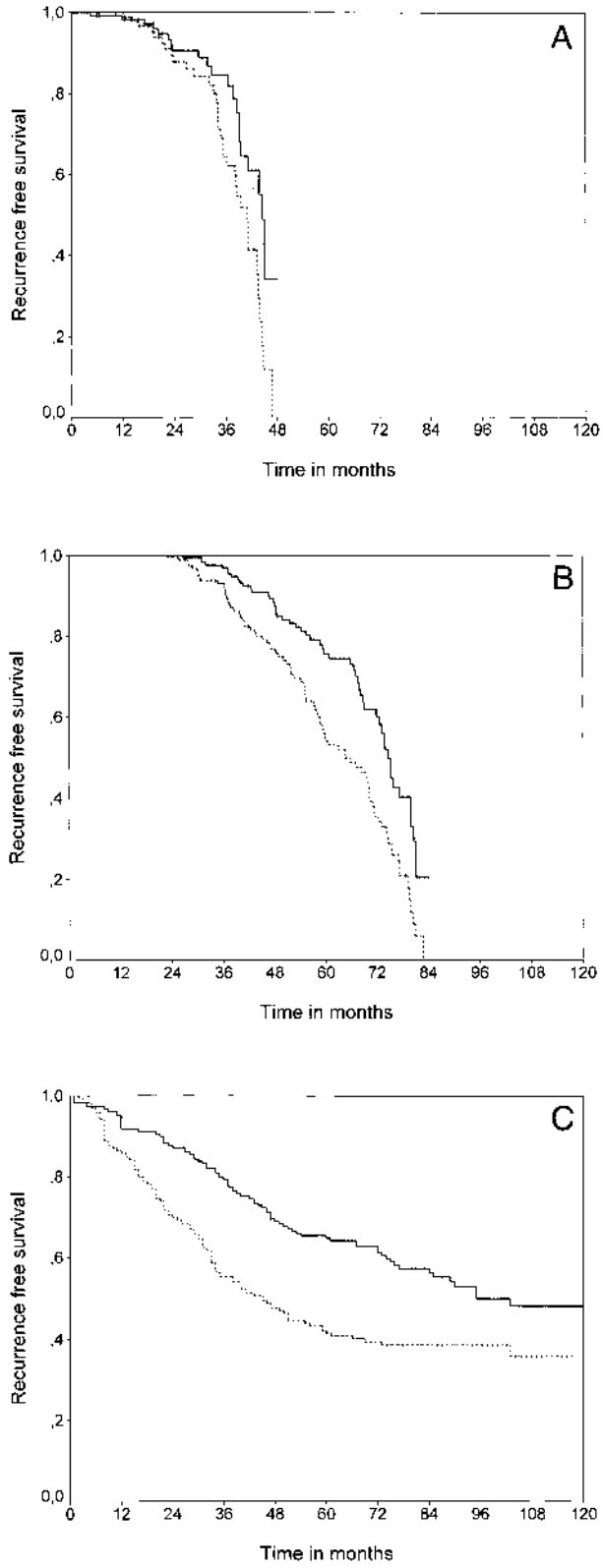

Fig. 3. Estimated survival functions by treatment group with different length of follow-up. The solid curve represents TAM and the dotted curve control treatment. A) follow-up 4 years from study start (266 patients included), B) maximum follow-up 7 years (350 patients), and C) 10 years (350 patients).

If the expected time to an event (e.g. death or recurrence) is short compared with the time needed to enter patients in a trial, a group sequential trial will on average require fewer patients than a fixed sample trial. If the time to event were long, however, the expected reduction in patient numbers would be negligible. Still, it might prove valuable to carry out interim analyses since the expected follow-up time required to reach a conclusion would be shorter. For the benefit of future patients, treatment policies might then be altered at an earlier stage. With long-term treatment even patients included in the trial might benefit from changing to the better therapy.

Both group sequential methods presented in this report maintain the desired overall significance level equally well. O'Brien and Fleming's method has the advantage of maintaining the desired power, whereas the method proposed by Pocock results in a loss of power with an increasing maximum number of tests. As a consequence, to maintain the desired power of a trial, the maximum sample size to be accrued must be increased by 10–20%, depending on the maximum number of analyses planned. The maximum number of patients that may be required will thus be greater with Pocock's method. With the O'Brien and Fleming method there is no need to increase the sample size; the sample size estimated for a fixed sample trial will suffice.

Another drawback of the Pocock stopping rules is that relatively more weight is placed on early analyses, when less information has accrued. It is thus possible that a fairly low p-value will be observed at the final analysis, but that the result will not be judged significant by the group sequential test. Many people find this confusing. The O'Brien and Fleming method puts more weight on later analyses, and the last analysis is undertaken at a nominal level, which is not very much lower than 0.05.

The choice of method depends on which properties are considered important in a specific trial. A possible advantage of applying the method developed by Pocock is that the average number of patients included at termination will be lower than that with the O'Brien and Fleming method. If it were important to stop as early as possible, Pocock's method would thus seem to be a reasonable choice. However, one would need to increase the maximum number of patients or events in order to maintain the desired power. If one were not prepared to increase the maximum size or length of the trial, the O'Brien and Fleming method would be the natural choice. The probability of stopping very early is lower than with Pocock's method, but the average number of events would be much lower than with a fixed size trial.

When comparing survival data, it is not necessarily wise to stop a trial very early. There is always the risk that survival data are not mature at the time of analysis because the follow-up has been too short. Early separation of survival or time to recurrence curves cannot be regarded

Table 4

Results of different analysis strategies

	Maximum 3 analyses		One final analysis Fixed sample
	Pocock	O'Brien & Fleming	
Years to termination	4	7	10
No. of patients	266	350	350
No. of events			
TAM	22	53	79
Control	34	88	104
Median time to recurrence (months)			
TAM[1]	45 (40, 48)	75 (70, 79)	95 (70, [2])
Control[1]	41 (38, 44)	64 (58, 70)	45 (34, 57)

[1] 95% confidence intervals in parentheses.
[2] No upper limit can be estimated.

as convincing evidence of a long-term remission or survival difference. Sometimes curves even cross, and the treatment that seemed to be worse during the first months may turn out to be the better in the long run. A too early termination of a trial might then lead to the wrong conclusion concerning treatment effect, and the crossing over of survival curves might not be discovered if the trial were stopped early. O'Brien and Fleming's method is thus generally considered to be the better choice. The possibility of stopping early is then extremely small unless treatment differences are very large.

A fundamental dilemma of early stopping of a clinical trial is to balance ethical considerations for individual patients in a particular trial versus the collective ethics of the community (9). Focus on individual ethics implies stopping randomization as soon as one treatment can be demonstrated to be preferable. The community would take the opposite view. Accurate information on the benefits and costs of treatment is of great value, and the community thus calls for larger trials with long-term follow-up.

The disadvantage of early stopping is that one might be left with an incomplete picture of the relative benefits of the treatments, due to short follow-up and less precise and possibly exaggerated estimates of treatment effect. Unduly enthusiastic recommendations often follow. On the other hand, small trials may not seem very convincing, and their impact on routine clinical practice may turn out to be negligible. Early termination will also reduce the possibility to perform important subgroup analyses because of the very low patient numbers in each subgroup.

The main reasons for stopping a trial early are a conlcusive result on the primary endpoint, either from the trial being conducted or from independent trials, or serious adverse effects on one of the treatments. The actual decision to stop a trial should not be based solely on the statistical stopping rules. These rules should perhaps instead be viewed as stopping guidelines. Current clinical knowledge, practical problems such as toxicity of treatments, and future research ideas must also be considered before deciding to terminate a trial early.

ACKNOWLEDGEMENTS

I thank the Norwegian Breast Cancer Group for providing data from the breast cancer study.

REFERENCES

1. McPherson K. Statistics: the problem of examining accumulating data more than once. N Engl J Med 1974; 290: 501–2.
2. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977; 64: 191–9.
3. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35: 549–56.
4. DeMets DL, Gail MH. Use of log-rank tests and group sequential methods at fixed calendar times. Biometrics 1985; 41: 1039–44.
5. Pocock SJ. Interim analyses for randomized clinical trials: the group sequential approach. Biometrics 1982; 38: 153–62.
6. Birtwistle GM, Dahl OJ, Myhrhaug B, Nygaard K. Simula begin. Bromley: Chartwell-Bratt Ltd, 1983.
7. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976; 34: 585–612.
8. Gundersen S, Hannisdal E, Soreide JA, Skarstein A, Varhaug JE. Adjuvant tamoxifen for pre- and postmenopausal women with estrogen receptor positive, node positive breast cancer: a randomized study. Breast Cancer Res Treat 1995; 36: 49–53.
9. Pocock SJ. When to stop a clinical trial. Br Med J 1992; 305: 235–40.
10. Fleming TR, Watelet LF. Approaches to monitoring clinical trials. J Natl Cancer Inst 1989; 81: 188–93.