Eur Respir J 2000; 15: 821–827
Printed in UK – all rights reserved

European Respiratory Journal
ISSN 0903-1936

Sequential re-analysis of a phase-III clinical trial in non-small cell lung cancer

N. Donaldson*, R.O. Dillman**, J. Wallace***, A. Ortiz-Hurtado+

Sequential re-analysis of a phase-III clinical trial in non-small cell lung cancer. N. Donaldson, R.O. Dillman, J. Wallace, A. Ortiz-Hurtado. ©ERS Journals Ltd 2000.

ABSTRACT: This paper presents a reanalysis of a randomized clinical trial conducted by the Cancer and Leukemia Group B (CALGB, Bethesda, MD, USA). This trial found a significant benefit of combination chemotherapy followed by irradiation (CTRT) in comparison to radiotherapy alone (RT) for the treatment of nonsmall cell lung cancer. The validity of the results obtained and the decision to terminate taken by the CALGB, were assessed using sequential methods. The reliability and efficiency of sequential methods were also assessed for this study.

Two sequential designs were used: the triangular and the restricted procedure. Initial analyses were conducted with the data from patients actually recruited, adjusting for important prognostic variables at any interim analysis. As a confirmatory technique, a continuation of the trial was simulated, sampling extra patients under the assumption of no treatment difference, preserving the effect of the prognostic variables.

Using the results from the 155 patients recruited by the CALGB (88 deaths at termination and 136 after follow-up), the sample path stayed within the continuation region of both sequential designs considered. An underpowered sequential analysis showed significant superiority of CTRT over RT (95% confidence interval (95% CI) 0.50–0.96, p=0.03 for the triangular; 95% CI 0.37–0.88, p=0.01 for the restricted procedure). Conventional analysis of the follow-up data also showed significant superiority of CTRT. The trial extended with simulated data ended at 60 months with 251 patients (178 deaths), showing significant superiority of CTRT under both designs (95% CI for hazard ratio 0.55–0.97).

The two sequential procedures would have led to the same conclusion as that reached by the Cancer and Leukemia Group B, still achieving considerable savings in patients recruited and time over the conventional design. The data simulated under the rather conservative null hypothesis did not reverse the positive result claimed by the Cancer and Leukemia Group B.

Eur Respir J 2000; 15: 821–827.

*Dept of Statistics, Stanford University, Stanford, CA, USA. **Hoag Cancer Center, Newport Beach, CA, USA. ***Dept of Physics, Astronomy and Mathematics, University of Central Lancashire, Preston, UK. +Datalab Statistical Consulting & Training, London, UK.

Correspondence: N. Donaldson
Dept of Mathematics
Imperial College
180 Queens Gate
London
UK
Fax: 44 2073463208

Keywords: Fixed-sample analysis
non-small cell lung cancer
O'Brien-Fleming design
randomized clinical trial
sequential analysis
triangular design

Received: May 20 1999
Accepted after revision December 31 1999

A great part of this work was carried out under Grant No. G9008019 from the British Medical Research Council.

The aim of the present study was to reanalyse data from a trial conducted by the Cancer and Leukemia Group B (CALGB), Bethesda, MD, USA [1] using sequential methods [2]. The trial (CALGB-8433) was designed to compare two therapies in the treatment of stage III non-small cell lung cancer, in terms of overall survival. The experimental treatment was combination chemotherapy followed by irradiation (CTRT) and the standard treatment was radiotherapy alone (RT). The trial was a prospective randomized nonblinded study. The CALGB conducted semiannual analyses of the data and terminated the study early in response to a treatment difference emerging over time. A truncated O'Brien-Fleming boundary [3] to account for this multiple testing was used to make the decision to terminate the study. The trial was judged to have been terminated prematurely [4] and was presented as a special case study of group sequential stopping rules [5]. The present study has two objectives: firstly, to assess the results obtained by the CALGB and, second, to compare the reliability and efficiency of two widely used boundary-based sequential methods with those of conventional procedures. In order to accomplish these objectives, a series of interim assessments were performed on the data generated by the trial and, if the amount of information (*i.e.* the number of deaths accumulated) was insufficient, supplemented with data simulated under the hypothesis of no treatment difference.

For editorials comments see page 819–820.

Patients and methods

Patients

All the patients had documented non-small cell cancer of the lung, including squamous cell carcinoma, adenocarcinoma and large cell anaplastic carcinoma. In addition, patients had to have stage III disease, established by clinical or surgical staging. This cancer has been considered

incurable because of the association of the locally extensive or invasive disease and the involvement of mediastinal lymph nodes in micrometastatic disease. The eligibility criteria included excellent performance status (*i.e.* able to perform normal activities or restricted only in vigorous activity), minimal weight loss (<5% during the preceding 3 months), visible disease on radiography and that the patient had not received chemotherapy or radiation therapy for this cancer. Randomization was conducted using a categorization by histological type followed by completely random allocation, to minimize the imbalance of this prognostic factor in the treatment groups. Patients in the CTRT group received vinblastine on days 1, 8, 15, 22 and 29, and cisplatin (100 $mg \cdot m^2$ body surface $area^{-1}$ on days 1 and 29). They then began a 6-week period of radiation therapy on day 50. Patients in the RT group began the same radiation therapy within 5 days of randomization and did not receive any chemotherapy. The main objective of the study was to compare overall survival in the two treatment groups. Clinicians assessed the response 1 month after the last dose of treatment and 2-monthly thereafter. Complete response was defined as the complete absence of measurable disease for ≥4 consecutive weeks [1]. The data that was available corresponded to the 155 patients recruited by the CALGB during a period of 34 months, updated to include follow-up data up to 1992.

Study design

The Cancer and Leukemia Group B design. The CALGB designed the trial to have an 80% probability of detecting a 50% change in survival after 2 yrs, at the 5% significance level. This means that the CALGB set out their power requirement to detect a difference in median survival from 9 months (in the RT group) to 13.5 months (in the CTRT group), or, equivalently, a hazard ratio (HR) of 0.666. In order to fulfil this requirement it was necessary to accumulate 191 deaths for fixed-sample analysis. The CALGB anticipated a recruitment rate of 20 $patients \cdot month^{-1}$ for 1 yr and that the final analysis would take place when 80% of the patients had died. Extrapolating the expected survival probabilities according to an exponential distribution, the CALGB expected this to occur after 3.5 yrs. A formal protocol-monitoring committee participated in each interim assessment of the trial, with the accumulating data analysed in detail twice a year. Group sequential methods were used at each interim assessment. The final decision to terminate the trial was based on the p-value of a truncated O′Brien-Fleming stopping rule.

Design of the sequential reanalysis. The first retrospective interim analysis was scheduled to take place on July 2, 1985, 1 yr after the first patient entered the trial. Subsequent interim analyses were scheduled at 6-monthly intervals for the first 2 yrs and yearly thereafter, to coincide with the protocol-monitoring committee meetings when the CALGB interim analyses took place. Patients alive at the time of an interim inspection were regarded as censored observations at that time. (A survival time is called censored if its precise value is not known; it is only known to continue beyond the time at which the interim analysis took place.) With sequential designs, recruitment has to remain open until termination of the trial. In consequence, an interim inspection was planned to take place on April 13, 1987, when the CALGB enrolled the last patient. If no conclusion were reached then, a comparison that was less powerful than that intended and which still adjusted for the sequential nature of the trial would have been produced. This is known as underrunning analysis [5]. Then, a continuation of recruitment was stimulated until a stopping boundary was reached. In order to test the result obtained by the CALGB the present simulation was undertaken with the conservative hypothesis of no treatment difference.

The treatment difference was expressed as the HR [6]. At termination of the trial, sequential methods provide an adjustment of two biases, the bias introduced by multiple testing (adjusting the p-value) and the bias introduced by the overresponse brought about by the stopping rule (adjusting the estimator of the HR) [2]. After termination, the deaths of patients recruited towards the end continue accumulating. This follow-up data was incorporated into the analysis by replacing the last analysis with a new one at 72 months. This more powerful analysis was used to confirm the previous results [5].

Choice of a sequential design. In order to determine the type of power requirement (see *Power requirements in a sequential design* (*Appendix*)) needed for the reanalysis, aspects of the cost and toxicity of the experimental treatment (CTRT) in relation to the control (RT) were taken into account. As for any cancer that was incurable by surgery, radiation therapy had been the treatment of choice for non-small cell lung cancer, and adding chemotherapy was probably seen as increasing the risk of toxicity. For this reason and because of the added cost, it was assumed that a clinician would have continued using RT in the case of equivalence. On this basis, an asymmetric power requirement seemed appropriate. Assuming that it was desirable to stop the trial early in the case of no treatment difference, a triangular design for the reanalysis was considered [2]. The power for detecting superiority of CTRT (over RT) was chosen to be 80%, whereas the power for detecting its inferiority was reduced to only 23%.

For completeness, a sequential design with a symmetric power requirement and a large expected sample size in the case of no treatment difference was also considered. The restricted procedure with a slope of 0.10, a quarter of the log HR, was chosen. This test was set to have a power of 80% for detecting superiority or inferiority of CTRT in relation to RT. The choice of the restricted procedure has the advantage that it relates closely to the stopping rule that the CALGB used as the O′Brien-Fleming design used by them is a restricted procedure with zero slope [5].

The accumulating evidence of the advantage of CTRT was summarized, in terms of the log-rank statistic [6]. This statistic was denoted by Z and adjusted for prognostic variables. As the interim analyses proceeded, a sample path was obtained by plotting Z against the amount of information accumulated (sample size), denoted by V. This statistic V is approximately equal to one quarter of the number of deaths [6]. The straight-line boundaries that determined stopping in this study (a triangle for the triangular design and a trapezoid for the restricted procedure) are shown in figures 1 and 2.

Expected sample size. The expected number of deaths termination is shown in table 1, for hypothetical values of

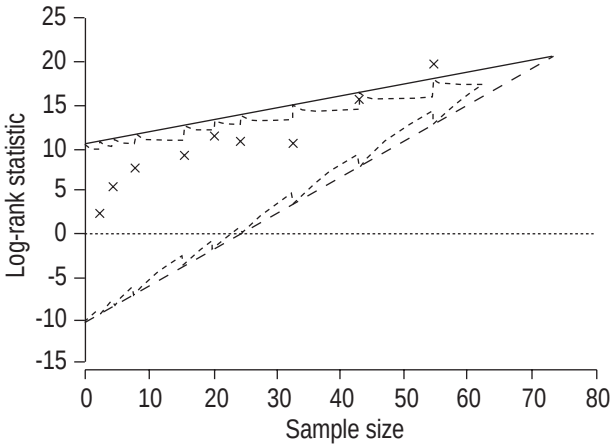

Fig. 1. – Triangular design for the Cancer and Leukemia Group B-8433 trial, with 80% power and 5% significance level. Sample size: a quarter of the number of deaths.

the treatment difference. In all cases, the expected number of deaths at termination is well below 191, the number required in a conventional design. In contrast, the maximum sample size is larger, but with an associated low probability. This is typical of a sequential design (see *Sample size at termination in a sequential design* (*Appendix*)).

Computer software

The computer package Statistical Analysis System (SAS; SAS Institute, Cary, NC, USA) was used for the nonsequential statistical calculations; in particular, procedure Proportional Hazard Regression (PROC PHREG) was used for survival analysis. Simulations programs were written in Fortran77. Finally, derivation of the stopping boundaries and adjustments to account for the sequential analyses were obtained using the computer package PEST3 developed by the Planning and Evaluation of Sequential Trials project at the University of Reading [7].

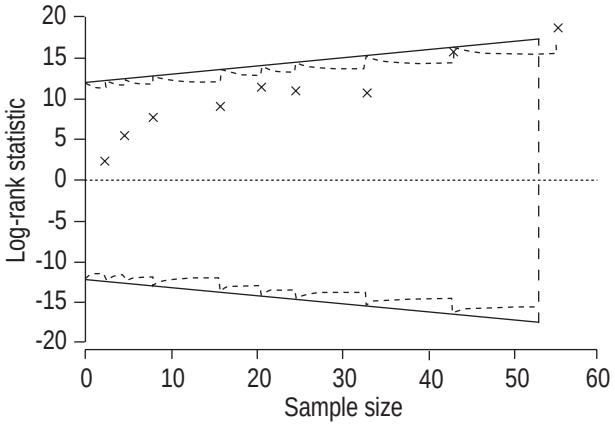

Fig. 2. – Restricted Procedure for the Cancer and Leukemia Group B-8433 trial, with 80% power and 5% significance level. Sample size: a quarter of the number of deaths.

Table 1. – Expected number of deaths at termination*

	Triangular	Restricted Procedure	O′Brien-Fleming
Upper boundary	Z=10.5+0.14V	Z=12+0.10V	Z=15.9
Max No. Deaths	295	212	202
Hazard ratio			
1.5 CTRT INF	51	142	147
1.22 CTRT (mod) INF	67	193	188
1 No treatment diff	96	209	200
0.82 CTRT (mod) sup	133	193	188
0.66 CTRT sup	127	142	147

*: 1 -β=0.80; α=0.05. Max: maximum; Z: log-rank statistic; V: sample size: a quarter of the number of deaths; CTRT: combination chemotherapy followed by irradiation; INF: inferior; mod: modestly; sup: superior; diff: difference.

Results

The Cancer and Leukemia Group B results

The CALGB enrolled patients from July 2, 1984 to April 13, 1987 at the unexpectedly low rate of approximately four per month. Interim analyses of survival times took place for March 1986, August 1986, October 1986 and March 1987. Several stopping rules were considered, but the trial was terminated using the O′Brien-Fleming boundary [2], truncated at 3.0 SD. At the time of stopping, 155 eligible patients had been enrolled but there were follow-up data for only 105 patients. A summary of the number of patients with follow-up data available and the number of deaths obtained at each interim analysis, categorized by treatment group, is presented in table 2. The log-rank statistics and associated p-values together with the significance levels of the truncated O′Brien-Fleming boundary are also shown. The log-rank p-values were unadjusted for the effect of prognostic variables [1].

A formal protocol-monitoring committee participated in the interim assessments of the trial. Throughout the period of the study, the CTRT group appeared to have the better survival outcome, but the study was continued, as the observed p-value did not cross the truncated O′Brien-Fleming boundary. At the March 1987 interim analysis, approximately 34 months after the first patient entry, the Cox′s regression model was used to adjust for sex, age, cell type and performance status, giving a p-value of 0.0008. This adjusted p-value fell below the boundary significance level, and the trial was stopped with rejection of the null

Table 2. – Interim analyses performed by the Cancer and Leukemia Group B

	No. of patients		No. of deaths		p-value	
	RT	CTRT	RT	CTRT	Log-rank	O′Brien-Fleming
March 1986	38	41	4	12	0.021	0.001
August 1986	47	41	14	20	0.0071	0.001
March 1987	54	51	24	32	0.0015	0.001

RT: radiotherapy alone; CTRT: combination chemotherapy followed by irradiation.

Table 3. – Observed survival patterns in the Cancer and Leukemia Group B-8433 trial

Inspection		No. of patients		No. of deaths		Unadjusted		Adjusted	
Date	Time point months	CTRT	RT	CTRT	RT	V	Z	V	Z
July 2, 1985	12	21	22	7	3	2.42	2.77	2.41	2.55
January 2, 1986	18	35	36	14	6	4.62	5.98	4.66	5.67
July 2, 1986	24	55	53	22	16	8.23	7.29	7.98	7.85
January 2, 1987	30	68	67	37	23	16.27	8.40	15.69	9.20
April 13, 1987	34	77	78	48	40	20.87	11.46	20.50	11.60
Simulated recruitment starts									
July 2, 1987	36	83	81	55	48	24.76	11.38	24.50	11.11
July 2, 1988	42	104	104	74	64	33.63	12.99	32.69	10.86
July 2, 1989	60	125	126	97	81	43.49	19.25	42.91	16.01
July 2, 1990	72	156	143	121	99	53.08	24.19	52.49	21.54
July 2, 1991	84	178	172	151	122	66.26	30.56	65.69	24.33

hypothesis of no treatment difference. The overall differences in survival were especially marked for patients with adenocarcinoma (p=0.016) and squamous cell carcinoma (p=0.0085) [1].

Various response variables were analysed by the CALGB: recurrence-free time reached marginal significance (p=0.04) and the rate of response to treatment was found to be marginally nonsignificant (p=0.09). The level of toxicity observed in the CTRT group caused serious concern. The sequential analysis presented here is for survival times only, the principal endpoint of the CALGB-8433 trial. The secondary endpoint, recurrence-free time, obviously correlated with survival times and its analysis was affected by the sequential stopping mle. As a consequence, the above p-value could have been more significant. A conventional analysis of secondary endpoints is not considered in this paper.

Results of the sequential reanalysis

In table 3, a summary of the data accumulated at the various interim analyses and the Z and V statistics obtained throughout the period of the study are presented. Cox′s regression model was used to investigate the relationship between survival times and possible prognostic characteristics at each interim analysis. None of the prognostic variables showed significance at the first (12-month) interim inspection. The Eastern Cooperative Oncology Group performance status was the only variable that showed significance between the 18-month and 36-month inspections. Cell type became significant, at the 10% level, at the 48-month inspection. The Z and V statistics were stratified accordingly. In table 3, summaries incorporating the further deaths accumulated after recruitment was closed are also shown.

The first interim analysis took place with 43 patients recruited and 10 deaths accumulated. The fifth inspection was carried out at 34 months, immediately after the CALGB closed recruitment (the last patient was recruited on April 13, 1987). Figure 1 shows the plot of the Z and V statistics. The mathematical derivation of the straight-line boundaries assumes continuous monitoring: the dotted "pine tree" boundary makes stopping easier, to compensate for the time intervals between inspections [2]. The sample path started increasing from the first inspection but had not crossed the upper boundary by the fifth inspection (with 155 patients recruited and 88 deaths). This confirms that the trial was terminated prematurely. Nevertheless, as seen in table 4, underrunning analysis [2] showed evidence of significant superiority of CTRT over RT.

When the Z and V statistics obtained from the 72-month follow-up of these patients are plotted, the sample path remains in the continuation region, confirming once more that the trial was terminated prematurely. Again, underrunning analysis showed evidence of superiority of CTRT (p=0.03). For comparison, conventional fixed sample analysis using these follow-up data is presented in table 4. This led to rejection of the null hypothesis, in favour of CTRT (p=0.03).

Simulated continuation of the trial

The continuation of the trial was simulated by sampling patients at the rate of four patients per month, consistent with the observed rate. The simulation of death times was conducted under the conservative assumption of no treatment difference. The response was generated randomly from an exponential survival distribution fitted to the times

Table 4. – Sequential analysis at 34 months*

	Time point months	No. of deaths	HR	95% CI for HR	p-value
Standard Triangular	34	88	0.69	0.50–0.96	0.03
Follow-up Triangular	72	136	0.68	0.48–0.97	0.03
Standard Restricted Procedure	34	88	0.57	0.37–0.88	0.01
Follow-up Restricted Procedure	72	136	0.68	0.48–0.97	0.03
Follow-up conventional (nonsequential)	72	136	0.68	0.48–0.97	0.03

*: 155 patients. HR: hazard ratio; CI: confidence interval.

Table 5. – Sequential analysis of the simulated continuation of the trial*

	Time point months	No. of deaths	HR	95% CI for HR	p-value
Standard Triangular	60	178	0.70	0.58–0.97	0.03
Follow-up Triangular	72	232	0.71	0.54–0.96	0.03
Standard Restricted Procedure	60	178	0.70	0.52–0.96	0.03
Follow-up Restricted Procedure	72	232	0.71	0.54–0.95	0.02
Follow-up conventional (nonsequential)	72	232	0.69	0.53–0.91	0.01

*: 251 patients; HR: hazard ratio; CI: confidence interval.

observed in the corresponding level of the prognostic variables. This was performed for the newly recruited patients, as well as for the patients that were still alive at the beginning of the simulation. The simulated Z and V statistics were plotted at the end of the sample path shown in figure 1. Under both sequential designs the sample path crossed the upper boundary at the 60-month inspection with 251 patients recruited and 178 deaths accumulated. The null hypothesis was rejected in favour of CTRT (p=0.03). These results are shown in table 5.

Incorporation of the 72-month follow-up data from these patients led to the rejection of the null hypothesis in favour of CTRT (p=0.03). For comparison, conventional fixed-sample analysis adding the follow-up and simulated follow-up data, is presented in table 5. A highly significant effect of CTRT was found (p=0.007).

Discussion

The CALGB investigators chose on ethical grounds not to continue randomizing patients to a treatment that was showing itself, consistently to be inferior. As a consequence of this they terminated the trial with 56 deaths approximately a quarter of the number required for the conventional method. On this basis the trial was judged to be underpowered and this was confirmed by the present sequential analyses. Using the results from the 155 patients recruited by the CALGB (88 deaths at termination and 136 deaths after follow-up), the sample paths of our sequential designs stayed within the continuation regions. However, the confirmatory analyses pointed toward significant evidence of superiority of CTRT over RT. In particular, the simulation of additional patients recruited under the hypothesis of no treatment difference failed to reverse the positive result shown by the underrunning analysis.

Generally the potential saving in sample size expected in a sequential design must be set against the slight chance of having to accumulate a sample larger than that required by the conventional design. Nevertheless this is an additional positive feature of sequential methods, since in the presence of null or moderate treatment differences these methods yield large samples, which allow more accurate estimation of treatment effect and toxicity. However, the maximum sample size in a sequential design is rarely attained, and, in fact, the two sequential procedures considered in this trial (triangular and restricted procedure) would have led to the same conclusion of superiority of CTRT, with considerable savings (patients recruited, deaths accumulated and duration of trial) over the conventional fixed-sample design.

The present results can be used to make observations regarding the relative efficiency of the triangular and restricted procedure designs, in terms of the amount of information required for completion of a trial. In general, the expected sample size at termination was smaller for the triangular design. In the case of no real treatment difference, the triangular test would result in rapid termination. If no early stopping was necessary in such a case, a restricted procedure should be chosen. This would be the case when monitoring safety and evaluating secondary effects is important.

The O′Brien-Fleming rule, as any α-spending function, provides a way of adjusting the significance level for the multiple looks, but it does not adjust for the bias introduced into the estimation by the sequential nature of the trial. This last feature is an advantage of the triangular and restricted procedure over the O′Brien-Fleming stopping rule. Estimation using the conventional techniques used in a fixed-sample analysis should not be conducted after a sequential trial is terminated (see *Approaches to sequential methods (Appendix)*).

All of the present results are based on analyses adjusted for important prognostic variables (see *Adjustment for Baseline characteristics (Appendix)*). For comparison, unadjusted and adjusted values are presented in table 3. In recruiting patients for the trial, the CALGB used a stratified randomization procedure by histological type, but not in terms of other variables which proved to be important, such as performance status. V, the amount of information contained in the data about the treatment effect, decreased as a result of stratification. However, the value of Z also changed such that the ratio Z:V increased. In consequence, the adjusted sample path moved towards the upper boundary at a quicker pace than the unadjusted one. The adjustment started making a distinctive impact on the calculated statistics V and Z from approximately the fourth interim analysis, at 30 months with 60 deaths accumulated. Following the rule proposed by the current authors (see *Adjustment for Baseline characteristics (Appendix)*), this is when the current authors would have scheduled the first interim analysis for monitoring this trial.

Sequential designs offer clinicians a scientific method for reacting to the continuous flow of evidence, allowing them to reach valid conclusions with considerable savings of time and resources. Sequential designs have been shown to protect from inconsistent results at termination and after follow-up is incorporated. That is, once a test has been stopped with a significant p-value, it is unlikely (although not impossible) that a later analysis which incorporates overrunning will be nonsignificant. This has been shown in the context of the triangular design [8], and, more

recently, with more extensive and systematic evidence, for the O′Brien-Fleming rule [9]. Researchers in the USA have used repeated significance testing widely with the O′Brien-Fleming rule the most popular method. Sequential methods based on the boundary approach have now made an impact on how phase III clinical trials are conducted, in the pharmaceutical and public sector in Europe and the UK [10]. It appears to the present authors that what left the CALGB-8433 trial open to controversy was the fact that a formal stopping rule had not been completely specified at the design stage of the trial. It is important to formulate a formal stopping rule at the outset, incorporating adjustment for prognostic variables and appropriate methods of estimation, which adjust for the possible bias arising from early termination.

From a practical viewpoint the most important aspect of the Cancer and Leukemia Group B-8433 trial is the impact it has had on the standard practice of medicine in the management of regionally advanced non-small cell lung cancer. Although a variety of analyses have confirmed the statistical validity of the study the medical community has insisted on confirmatory documentation In response the long-term results of the Cancer and Leukemia Group B-8433 trial were published in the *Journal of the National Cancer Institute* in 1996, and continued to show the benefit of combination chemotherapy followed by irradiation relative to radiotherapy alone [11]. The initial report of the Eastern Cooperative Oncology Group/Radiation Therapy Oncology Group trial which was designed to confirm the Cancer and Leukemia Group B-8433 trial, suggested similar results after 1 yr of follow-up [12]. Meta-analysis of all randomized trials involving cisplatin-based chemotherapy for regionally advanced non-small cell lung cancer has confirmed the value of the addition of chemotherapy compared to radiation alone [13]. National Cancer Center Network guidelines and others, also now routinely recommend chemotherapy as part of the management of these patients. However, the impact of newer chemotherapeutic agents and whether there is enough advantage to concurrent chemotherapy and radiation therapy as compared with sequential chemotherapy followed by radiation therapy to offset the increased toxicity associated with concurrent use of chemotherapy and radiation therapy still remains a major issue.

Appendix: sequential methods in comparative experiments

The rationale

When a comparative experiment is conducted sequentially, multiple looks or interim analyses are performed with the intention of stopping recruitment of patients as soon as a significant treatment difference is detected, avoiding randomization of patients to a treatment known to be inferior. When this approach is taken, there are two immediate problems to be faced. The first problem is that the significance level (risk of type I error) is inflated by the multiple tests that are being performed in the same experiment [14]. The second problem relates to the bias (in favour of the treatment that emerges as superior) introduced by the fact that the trial is stopped precisely when a significant advantage of one of the treatments is observed. Sequential analysis is the branch of statistics that deals with appropriate solutions to these two problems.

Approaches to sequential methods

Stopping rules tackle the first problem, spreading the proposed significance level out (usually 5 or 1%) throughout the period of study. This may be done by specifying the proportion of the overall significance level used up until a given time. This is given in the form of a function, called the α-spending function, which increases from 0 to α, indicating that at termination all of the proposed significance level has been used up. An example of this is the O′Brien-Fleming stopping rule: its α-spending function rises very slowly when there is little information, making early termination difficult and rises quickly towards the end. An α-spending function can also be specified in terms of stopping boundaries, indicating what values of the observed treatment effect indicate superiority or inferiority of the experimental treatment or equivalence of treatments. Examples of these boundary-based sequential designs are the triangular and the restricted procedure designs used in the reanalysis presented in this paper (fig. 1). The restricted procedure with a slope of 0 corresponds to the α-spending function of the O′Brien-Fleming stopping rule.

To tackle the second problem, the class of boundary-based sequential designs provides a method of adjusting for the bias introduced by early termination. This adjustment is a function of the number and timing of the interim inspections that either take place before the trial is stopped or would have taken place had the trial not been stopped.

Power requirement in a sequential design

The sequential designs considered are derived in terms of either a symmetric or an asymmetric power requirement. An asymmetric power requirement guarantees a high probability of detecting superiority of the experimental treatment but it does not place any restriction on the probability of detecting inferiority. This type of power requirement is appropriate if, in the case of equivalence, the obvious choice for the clinician is the standard treatment. An example of this is the triangular design used in the reanalysis of the Cancer and Leukemia Group B (CALGB) trial (see *figure 1*). A symmetric power requirement guarantees a high probability of detecting superiority as well as a high probability of detecting inferiority of the experimental treatment relative to the control. In this case, it is necessary to acquire more information in order to allow the investigators to make a clear-cut distinction between equivalence and inferiority of the experimental treatment. A symmetric power requirement yields a large sample size in the case that no treatment difference exists. An example of this is the restricted procedure used in one of the reanalyses of the CALGB trial (see *figure 2*).

Sample size at termination in a sequential design

The size of the control and experimental groups in a clinical trial that is analysed sequentially can only be determined at the time the trial is terminated. Since it cannot be determined in advance of the experiment being conducted, the sample size is a random variable with a probability distribution, an expected value and a variability. The power requirement and the rate at which the significance level is used up with time dictate this distribution. It is good statistical practice to present the expected sample size at termination, under different scenarios of the real treatment

difference (see *table 1* for the sequential designs presented in this paper).

The expected sample size of a sequential clinical trial is very likely to be much smaller than the sample size dictated by the power requirement of the conventional fixed-sample design. However, the mathematical derivation brings up a small chance that it is larger. Fortunately, this larger sample size occurs only in the presence of a null or moderate treatment difference, and then only when a more accurate estimation (of both treatment effect and toxicity) is needed. It is also seen as a small price to pay for the possibility of an early termination in the presence of important differences. The maximum sample size associated with each sequential procedure should also be stated at the design stage (see *table 1* for the sequential designs presented in this paper).

There is considerable variation in the expected sample size at termination among the different sequential designs. For example, it is extremely unlikely that stopping will occur early under the O′Brien-Fleming boundary and the number of deaths at its termination is expected to be close to the number of deaths required in a fixed sample design. The greater the slope of the upper boundary in a restricted procedure, the easier it is for the trial to be stopped early. The present choice of slope was motivated by a compromise between the O′Brien-Fleming and Armitage′s boundaries [15]. The latter, with a boundary equal to half the log hazard ratio (0.20 in the present case), tends to allow early termination of the trial too often, even in cases in which there is not a genuine treatment difference [2].

The sample path

The accumulating evidence concerning treatment difference is summarized in terms of two statistics measuring treatment effect and sample size. The first statistic, denoted Z, provides a cumulative measure of the advantage of the experimental treatment over the control. In the present trial, Z is the log-rank statistic, adjusted for prognostic variables. The second statistic, denoted V, measures the amount of information, and, in the context of the present trial, is approximately equal to a quarter of the number of deaths. As the interim analyses proceed, Z is plotted against V [5]. The straight-line boundaries that determine stopping for this trial, a triangle for the triangular design and a trapezoid for the restricted procedure, are mathematically derived assuming continuous monitoring (see *figure 1*).

Adjustment for baseline characteristics

As performed in the present reanalysis, either stratification or covariate adjustment should be used at each interim analysis in a sequential design. The main reason for this is that randomization may fail to naturally balance the groups for important prognostic factors, in the case of the small samples that are likely to arise from an early termination. Although stratified randomization has been devised to correct for such imbalances, it may fail to balance the groups for a prognostic factor that has not been considered. However, these small samples will, most probably, fail to detect effects of prognostic variables that are clinically important. To this effect, the authors propose that the first interim analysis be scheduled after ~20% of the maximum sample size has accumulated. Simulations and reanalyses of different clinical trials suggest that such a rule guarantees enough power to detect effects consistently, from at least the second interim analysis until the end of the trial. Another important reason for adjustment is that stratification does not assume proportionality of hazards between strata and has been seen to perform well when the assumption of proportional hazards is not satisfied.

Acknowledgements. The authors would like to thank the Cancer and Leukemia Group B (CALGB), Bethesda, MD, USA for making the data available to them and the Dept of Applied Statistics at the University of Reading and the Medical Research Council Cancer Trials Office in Cambridge, UK for their interest in and cooperation with this research.

References

1. Dillman RO, Seagren SL, Propert KL, *et al*. A randomized trial of induction chemotherapy plus high-dose radiation *versus* radiation alone in stage III non-small-cell lung cancer. *New Engl J Med* 1990; 323: 940–945.
2. O′Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; 35: 549–556.
3. Souhami RL, Spiro SG, Cullen M. Chemotherapy and radiation therapy as compared with radiation therapy in stage III non-small cell cancer. *New Engl J Med* 1991; 324: 1136–1137.
4. Propert KJ, Kim K. Group sequential methods in multi-institutional cancer clinical trials. A case study. *In*: Peace KE, ed. Biopharmaceutical Sequential Statistical Applications. New York, Marcel Dekker, 1992; 133–153.
5. Whitehead J. Planning and Evaluation of Sequential Clinical Trials. 2nd Edn. Chichester, Ellis Horwood, 1992.
6. Parmar KB, Machin D. Survival Analysis. A Practical Approach. Chichester, Wiley, 1994.
7. Brunier H, Whitehead J. The PEST (Planning and Evaluation of Sequential Trials). Dept of Applied Statistics, University of Reading, Reading, UK, 1996.
8. Whitehead J. Overrunning and underrunning in sequential clinical trials. *Control Clin Trials* 1992; 13: 106–121.
9. Choi SC, Young JL. Interim analysis with delayed observations in clinical trials. *Stat Med* 1999; 18: 1297–1306.
10. Ritchie A, Oliver T, Fayers PM, *et al*. On the development of the MRC trial of α-interferon in metastatic renal carcinoma. *Stat Med* 1994; 13: 2249–2260.
11. Dillman RO, Herndon J, Segren SL, Eaton WL, Green MR. Improved survival after sequential chemotherapy-radiotherapy compared to radiation therapy alone in stage III non-small cell lung cancer: 7-year follow-up of CALGB 8433 trial. *J Nat Cancer Ins* 1996; 88: 1210–1215.
12. Sause WT, Scott C, Taylor S, *et al*. Radiation Therapy Oncology Group (RTOG) 88-08, and Eastern Cooperative Oncology Group (ECOG) 4588. Preliminary results of a phase III trial in regionally advanced unresectable non-small-cell lung cancer. *J Nat Cancer Ins* 1995; 87: 198–205.
13. Non-Small Cell Lung Cancer Collaborative Group. Chemotherapy and non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomized clinical trials. *BMJ* 1995; 311: 899–909.
14. Armitage P, McPherson K, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc A* 1969; 132: 235–244.
15. Armitage P. Sequential Medical Trials, 2nd Edn. Oxford, Blackwell, 1975.