



Trans Tasman Radiation Oncology
Group Limited
ACN 132 672 292

TROG

POLICY AND PROCEDURES

Statistical Guidelines

TPP E5

Version 3: 28 July 2009

(Always refer to the TROG website to check for the current version of this policy)

TROG Central Operations Office

Level 5, Building 7
Calvary Mater Newcastle
Locked Bag 7 HRMC NSW 2310
Tel: + 61 2 49 211 466
Fax: + 61 2 49 211 465
Email: trog@trog.com.au
Website: www.trog.com.au

Contents

1	Introduction	3
1.1	Purpose and Scope	3
1.2	The Role of Statistics in Clinical Trials	3
1.3	The Role of the Statistician in Clinical Trials.....	3
1.3.1	Protocol development	3
1.3.2	Trial conduct	3
1.3.3	Analysis and reporting.....	3
2	Trial Design	3
2.1	Phase II Trials.....	3
2.1.1	Single stage designs	3
2.1.2	Two-stage designs	3
2.1.3	Randomised phase II designs	3
2.2	Phase III trials	3
2.2.1	Principles of randomised trials.....	3
3	Objectives and Endpoints.....	3
4	Sample Size Estimation in a Phase III Trial	3
4.1	Phase III superiority trial with a survival-type endpoint	3
4.1.1	The elements of the sample size calculation.	3
4.1.2	Sample size checklist.....	3
4.1.3	Sample size adjustments	3
4.1.4	Sample size and a response rate outcome.....	3
4.1.5	Time to recurrence endpoints and competing events in sample size calculation	3
4.2	Phase III trials – Non-inferiority trial.....	3
4.3	Phase III trials sample size for a continuous outcome variable.....	3
5	Analysis and Interpretation of Trial Data	3
5.1	Interim analyses and premature termination	3
5.2	Group Sequential trials and Stopping Rules.....	3
5.3	Main Analysis and Publication.....	3
6	Miscellaneous topics	3
6.1	P-values and confidence intervals.....	3
6.2	Interpretation of negative trials	3
6.3	Small under-powered trials	3
6.4	The hazard ratio and the Cox proportional hazards model	3

6.5	Analysis of QoL data.....	3
6.6	Rules of Thumb.....	3
6.6.1	Confidence interval for a proportion when none (or all) patients respond	3
6.6.2	Approx sample size for comparing two binary proportions (e.g., response rates)	3
6.6.3	Relationship of P-values to outcomes	3
7	Acknowledgements	3
8	References.....	3

1 Introduction

1.1 Purpose and Scope

The following guide has been designed to serve as a guide for all participants in a clinical trial, but primarily for the clinician who is involved in the design of a phase II or III trial and who wishes to understand the statistical principles and requirements of the trial in order to facilitate working with the statistician in the most productive, efficient and scientifically valid manner.

This section also covers aspects of the conduct, analysis and interpretation of the trial. The approach is to provide principles and general guidance for the most common types of trials and situations that arise. A list of references provides further information.

In particular, the trialist is referred to national and international guidelines which, increasingly are required standards for the conduct of clinical trials.

1.2 The Role of Statistics in Clinical Trials

Statistics is the science of interpreting data that include random variability. Its practice includes accounting for how data were generated and assessing the validity and strength of conclusions that can be drawn from the data. The scientific validity of the trial will depend in part on the statistical soundness of the design, the conduct and the analysis and interpretation of the trial.

The primary instrument for acquiring valid evidence in clinical cancer research is the randomised clinical trial. The concept of randomisation in experimentation is essentially due to Sir Ronald Alwyn Fisher who promoted the technique primarily in agricultural trials (The Design of Experiments, 1935).

The first properly conducted randomised trial in medicine is said to have been that conducted by Sir Austin Bradford Hill in conjunction with the Medical Research Council in the UK; this trial, begun in 1946 (published 1947), compared bed rest with or without streptomycin in patients with tuberculosis.

The first randomised trial in cancer, begun in 1954 and conducted by the National Cancer Institute, USA, compared two regimens of 6-mercaptopurine and methotrexate in patients with acute lymphocytic leukaemia (Frei, Holland & Schneiderman, 1958).

The first randomised clinical trial in Australia was undertaken by the Australian Cancer Society; it compared cyclic with sequential drug treatment for children with

acute childhood leukaemia (ACS Childhood Leukaemia Study Group, 1968; Cox, 1968).

1.3 The Role of the Statistician in Clinical Trials

According to the ICH guideline Statistical Principles for Clinical Trials (E9) ,“the actual responsibility for all statistical work associated with clinical trials will lie with an appropriately qualified and experienced statistician, as indicated in ICH E6 (GCP Guidelines). The role and responsibility of the trial statistician, in collaboration with other clinical trial professionals, is to ensure that statistical principles are applied appropriately in clinical trials supporting drug development. Thus, the trial statistician should have a combination of education/training and experience sufficient to implement the principles articulated in this guidance.”

The trial statistician contributes to the design, the conduct, and the analysis and interpretation of a clinical trial. It is important that a dedicated trial statistician be involved from the inception of the trial but such collaboration will be enhanced if the Trial Chair is well grounded in the statistical principles behind clinical trials.

Of these three stages, perhaps the design phase is the one most important in ensuring that the statistical credentials of the trial are established; time spent in planning the trial at this stage is rarely misplaced. The statistician's role at this stage is, initially, in concept development and, later, development of the protocol, not only in providing the statistical considerations section (with the required sample size for the trial) but also in ensuring the Objectives and the Outcomes Measures sections are clear, consistent and comprehensive and as contributing to the overall quality of the protocol. Specifically, the main areas of a statistician's responsibility are:

1.3.1 Protocol development

To collaborate on:

- Concept development (initial and general discussion of appropriate designs and feasibility issues, such as main outcome criteria, approx. patient numbers required)
- Advice on trial design in relation to objectives;
- Delineation and definition of outcome criteria;
- Design of randomisation scheme including stratification factors;
- *Statistical Considerations* section of protocol; this should include the following items:
 - A description of the trial design and the trial's primary endpoint;

- A description of the method of allocation of patients to treatment arms;
- A description of the statistical methods that will be used, in particular the specific analysis to be used to address the primary objective of the trial. (This assists the statistician at the future date of final analysis and helps to avoid the dangers of 'data dredging' whereby multiple analyses are performed until one is found that yields a satisfactory result.)
- The sample size calculation, including a statement of anticipated accrual rate and projected termination date;
- The statistical analysis plan, including details of interim analyses and stopping rules and the expected timing of interim and main analyses during the conduct of the trial.

1.3.2 Trial conduct

- Serving on the Trial Management Committee
- Statistical aspects of protocol amendments to active trials.
- Interim reports for annual meetings: monitoring accrual, adverse events, overall trial performance on outcome measures;
- Interim analyses: for the Trial Management Committee and Independent Data Monitoring Committee;
- Participate in termination decisions.

1.3.3 Analysis and reporting

- Ensure a final Statistical Analysis Plan (SAP) is in place for the analysis of the trial data prior to commencement of the analysis. (The SAP may be included in the protocol or in a separate document.)
- Analyse the trial data according to the SAP and produce a Final Report addressing the main objectives of the trial; also analyse the data for secondary and other objectives and present results in a Report;
- Contribute to the production of manuscript(s) for publication and presentations.

2 Trial Design

The two main types of trials conducted by TROG are phase II (single-arm and randomised) and phase III trials. Other types of trials that may be conducted by TROG are other single-arm trials, pilot studies and exploratory studies, and prospective database studies.

2.1 Phase II Trials

Definition: a phase II trial is one conducted to determine whether a given therapy demonstrates sufficient efficacy to warrant its incorporation in a phase III trial. Usually other types of outcomes, such as safety and feasibility, will also be assessed.

Usually a phase II trial is a small-sample, single-arm trial but randomised phase II trials are also performed. Often the primary endpoint for a phase II trial is a response rate (a binary outcome). There are many different designs possible for a single-arm, phase II trial, ranging from a simple fixed sample size design to multi-stage designs (e.g. Simon two-stage) which have predetermined rules for stopping or continuing the trial at one or more pre-determined interim assessments.

Types of phase II trials

For single-arm phase II trials designs are usually either one-stage or two-stage (or more). In a one-stage design analysis is performed after all patients have been accrued. In a two-stage design interim analysis occurs after a specified number of patients have been accrued and their responses obtained and a decision is made to stop or continue depending on the results obtained. (Multi-stage designs are also possible).

In the following a number of designs require the specification of a minimum acceptable response rate (p_1) and a maximum unacceptable (p_0) response rate. A hypothesis test is then determined where, usually, $H_0: p=p_0$ versus $H_1: p=p_1$. (This is shorthand notation for $H_0: p \leq p_0$ versus $H_1: p > p_0$ (where power is determined for $p = p_1$.) One-sided testing is used and type I (α) and type II (β) errors are specified.

2.1.1 Single stage designs

Simple, single-arm, fixed size design

This is a trial usually of from 35 to 50 patients designed to obtain an estimate, with confidence interval, of the response rate. The standard error of the estimated response rate can be calculated as follows.

Binomial standard error and confidence interval (CI)

The standard error (se) of a proportion, p , (e.g. of a response rate) calculated from n patients is

$$se = \sqrt{p(1-p)/n}$$

For a p of around 50% (e.g. within the range, 30% to 70%) this is approximately,

$$se \approx 50/\sqrt{n} \text{ (percent)}$$

Ex.: the standard error of a response rate between 30% and 70% estimated from 100 patients is approximately, $se = 5\%$.

A 95% confidence interval for a response rate is approximately $p \pm 2 \times se$.

Ex. The observed response rate from 25 patients was 44%. An approximate 95% CI for the true response rate is $44\% \pm 20\%$; or 95% CI = 24% to 64%.

A form of simple hypothesis testing can also be done, as follows:

Example

Consider a trial of 40 patients, where a response rate of $p_0 = 10\%$ is the maximum unacceptable response rate and $p_1 = 0.30$ is the minimum acceptable rate. If the true response rate were 10%, we would see on average two patients achieve a response. According to Table 1 below, a result of 8 responses or more would be statistically significant at $P \leq .05$ (one-sided test) indicating that the true response rate is greater than 10%; a result of 6 responses or fewer would be statistically significant at $P \leq 0.05$ (one-sided test) indicating that the true response rate is less than 30%. The table can be used also to apply to adverse event rates.

Table 1. For a given sample size, N , and a response rate, p , either (i) a maximum unacceptable response rate or (ii) a minimum acceptable response rate the table cells provides values "a,b", where (i) "a" is the largest number of responses that is statistically significantly smaller than would be consistent with a true response rate of p , and (ii) "b" is the smallest number of responses that is statistically significant greater than would be consistent with a true response rate of p .

a,b	Response rate (<i>p</i>)											
N	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	
8	-,3	-,3	-,5	-,6	0,6	1,7	2,8	2,-	3,-	5,-	5,-	
10	-,3	-,4	-,5	0,6	1,8	1,9	2,9	4,10	5,-	6,-	7,-	
12	-,3	-,4	-,6	0,7	1,9	2,10	3,11	5,12	6,-	8,-	9,-	
15	-,3	-,5	0,7	1,9	2,10	3,12	5,13	6,14	8,15	10,-	12,-	
20	-,4	-,5	0,8	2,10	3,13	5,15	7,17	10,18	12,20	15,-	16,-	
25	-,4	-,6	1,9	3,12	5,15	7,18	10,20	13,22	16,24	19,-	21,-	
30	-,5	0,7	2,11	4,14	7,17	10,20	13,23	16,26	19,28	23,30	25,-	
35	-,5	0,8	2,12	5,16	8,20	12,23	15,27	19,30	23,33	27,35	30,-	
40	-,5	0,8	3,13	6,18	10,22	14,26	18,30	22,34	27,37	32,40	35,-	
45	-,6	0,9	4,15	8,20	12,24	16,29	21,33	25,37	30,41	36,45	39,-	
50	-,6	1,10	5,16	9,21	13,27	18,32	23,37	29,41	34,45	40,49	44,-	
75	0,8	2,13	8,22	15,30	22,38	29,46	37,53	45,60	53,67	62,73	67,75	
100	1,10	4,16	13,28	22,39	31,49	41,59	51,69	61,78	72,87	84,96	90,99	

Gehan design

A single arm, fixed sample size trial in which a minimum acceptable response rate (p_0) is specified and the sample size, n , is determined such that the drug will be deemed to have failed if the number of responses is less than a critical number, which event would occur with at most 5% chance if p_0 were the true response rate. Typically, $p_0 = 0.20$ and $n = 14$ and 0 responses out of 14 means failure of the drug and one or more responses means that investigation of the therapy is continued.

Fleming test

A single arm, single stage design in which a minimum acceptable (p_1) and a maximum unacceptable (p_0) probability are specified and a formal hypothesis testing situation is set up. One-sided testing is used.

Formal hypothesis testing can be performed by either of two approaches:

Pessimistic approach: $H_0: p \leq p_0$ $H_1: p > p_0$ (power for $p = p_1 > p_0$)

Optimistic approach: $H_0: p \geq p_1$ $H_1: p < p_1$ (power for $p = p_0 < p_1$)

where p_0 is the maximum unacceptable response rate and p_1 is the minimum acceptable response rate. The sample size can then be calculated once the type I error (α) and power ($1-\beta$) are specified.

Symmetric confidence interval design

This is the same as the Fleming test except that the type I and II errors are set to be the same (e.g. $\alpha = \beta = 0.10$). In this case the sample size is calculated such that an 80% CI, say, with width equal to $p_1 - p_0$, will either include p_0 and exclude p_1 or will exclude p_0 and include p_1 ; i.e. a decision is made for one or the other and no indeterminate result is possible (in theory).

2.1.2 Two-stage designs

Simon two-stage design

p_0 and p_1 and the hypotheses are as specified as above, with $H_0: p \leq p_0$. α and $1-\beta$ are the type I error and power, as before. The design determines a set of parameters (r_1, n_1, r, n) where n_1 is the sample size for stage 1 and r_1 the critical value, such that if the number of responses in the first n_1 patients is r_1 or less the trial stops and the drug is deemed to be unacceptable; and n is the overall sample size at the end of stage 2 and r is the critical value, such that if the number of responses is r or less from the n patients the drug is deemed to be unacceptable.

A number of parameter sets satisfy the above specifications, however two of these are considered more desirable. The *optimal* design is that which minimises the expected total sample size (given that p_0 is the true response rate) and the *minimax* design is that which minimises the maximum sample size, n (if stage 2 is undertaken). (See reference 14 for details.) The design details ($r_1, n_1; r, n$) have to be determined from tables or a computer program.

Ex.: If $p_0 = 0.20$ is the maximum unacceptable response rate and $p_1 = 40\%$ is the minimum acceptable response rate, and $\alpha = 5\%$ (i.e. if $p = p_0$, the drug will be accepted with probability 5%) power $(1-\beta) = 80\%$ (i.e. if $p = p_1$, the drug will be accepted with probability 80%) the *optimal* design is:

(3, 13; 12, 43), i.e. stage 1: $n_1 = 13$ patients; stop if no. of responses $x_1 \leq r_1 = 3$, otherwise continue to $n = 43$ patients and declare drug unacceptable if the no. of responses, $x \leq 12$, otherwise acceptable.

and the *minimax* design is:

(0, 6; 10, 33), i.e. stage 1: $n_1 = 6$ patients; stop if no. of responses $x_1 \leq r_1 = 0$, otherwise continue to $n = 33$ patients and declare drug unacceptable if no. of responses, $x \leq 10$, otherwise acceptable.

2.1.3 Randomised phase II designs

There are two situations in which a randomised phase II trial has been used: (i) where several new therapies are simultaneously assessed in the one trial; and (ii) where a new therapy is trialled for effectiveness in parallel with a control group representing standard or conventional treatment. In this latter case, the aim is not to provide a definitive comparison with control (which may be allocated to only a minority of the patients) but, as well as estimating the response rate of the new therapy, to estimate the difference between it and control and provide a confidence for the true difference. The evidence for efficacy from the estimate and CI can then be assessed along with other considerations (such as safety and feasibility) as to whether to continue its evaluation in further trials.

2.2 Phase III trials

Definition: a phase III trial is one conducted in order to compare directly the efficacies of two or more therapies. Usually this is a comparison between a new (*experimental* or *test*) therapy and a conventional or standard therapy for the disease in question.

Usually a phase III trial is a large-sample, two-arm trial with a sample size large enough to enable a definitive statement to be made about the relative efficacy of the two therapies. The main endpoint for the trial will usually be a time-to-event outcome such as survival duration, failure-free survival, or time to relapse. There a number of designs possible for a two-arm, phase III trial, the simplest being a parallel, fixed sample, open label trial. A phase III trial may be a superiority (difference) trial or, less commonly, a non-inferiority trial.

A *superiority trial* is one which aims to demonstrate that the new treatment is more effective than the standard therapy. A *non-inferiority trial* is one in which the aim is to demonstrate that the new treatment is 'at least as good as' the standard treatment; this type of trial is usually conducted to determine whether a new treatment which is less toxic or less costly than the standard treatment can be substituted for the standard treatment.

Two common variations on, or extensions to, the two-arm, fixed-sample-size parallel design are:

- Factorial design – two (or more) separate treatment questions are simultaneously addressed in the one trial. For example, both the question of which of two radiation dose schedules is more effective and the question of whether the addition of concurrent chemotherapy is effective can be assessed. Radiotherapy schedule and concurrent chemotherapy are two *factors* each at two levels (dose1, dose2; and no chemotherapy, chemotherapy) giving $4 = 2 \times 2$ combinations or treatment arms. The efficacy of radiotherapy schedule, say, is assessed by collapsing over the chemotherapy arms. A problem occurs if there is a significant *interaction* between radiotherapy schedule and chemotherapy; i.e. the effect of chemotherapy at the lower radiotherapy dose may be quite different from its effect at the higher radiotherapy dose. Analysis may then need to be performed on subgroups of patients.
- Group sequential design – planned interim analyses are performed with a view to stopping the trial early for efficacy (to detect a large difference early) or for futility (in order to stop the trial early if there is little chance of eventually demonstrating a treatment difference).

A group sequential design should be used in large trials, where feasible, to ensure that no more patients than necessary are subjected to an inferior treatment.

2.2.1 Principles of randomised trials

The basic principles of design are randomisation, stratification and replication (numbers of patients participating). These features are designed to ensure, respectively, elimination of bias, control of the effects of prognostic factors on outcome, and adequate power to demonstrate treatment effects.

Elimination of bias

When two treatments are compared any difference in efficacy observed may be due to one or more of (i) a true difference between the treatments, (ii) a difference due to the distribution of an important prognostic factor or factors favouring one arm over the other and (iii) chance. It is hardly ever possible to eliminate the contributions due to (ii) or (iii) completely, however, their influences can be controlled or minimised.

Bias is defined as a *systematic* tendency for factors associated with the design, conduct or analysis of a trial to favour one arm over the other (i.e. tend to inflate its influence on outcome relative to the other arm). Bias can be either conscious bias or, more commonly, unconscious bias on the part of investigators. There are a number of common sources of bias, and these can occur at any of the design, conduct, analysis, interpretation and publication stages. Some examples are:

- (i) selection bias (ascertainment bias) in which patients are assigned to arms such that there is a tendency for better prognosis patients to be assign to one arm rather than the other;
- (ii) exclusion bias in which patients with poor prognosis are excluded from analysis selectively from one arm, either of necessity (withdraw for the trial) or by decision (excluded in the analysis);
- (iii) differential evaluation bias in which patients in one arm are evaluated more exhaustively or more often;
- (iv) subjective evaluation bias in which outcome assessment is subjective and favours, whether consciously or unconsciously by the assessor, one arm over the other.

Randomisation

Selection bias is eliminated by strict application of random assignment of patients to arms. This implies, among other things, that the treatment to be assigned to each patient can't be predicted by those entering the patient on the trial. Randomisation *tends* to balance the distribution of prognostic factors between the arms, however, contrary to common belief, it does not prevent imbalance. A 'significant' imbalance of a prognostic factor demonstrated at the end of the trial does not mean the randomisation

mechanism had broken down. What is important is that randomisation ensures that any imbalance was not more likely to have favoured the experimental arm, say, over the control arm, or *vice versa* (hence the use of the term *systematic* in the definition of *bias* above).

A principle in the application of randomisation is that, where feasible, randomisation should occur as late as possible. This is to prevent the situation where, for example, there is a gap between randomisation and start of treatment and patients withdraw from the trial, decide to be treated differently from the treatment they were randomised to or develop disease progression. Such patients must be accounted for and be included in the main analysis according to the arm to which they were assigned (Intention To Treat principle) but would create noise in the comparison which would not necessarily be overcome (neutralised) by accruing extra patients.

Blinding

If the assessment of an outcome is subjective it is important to try to eliminate the subjectivity or, at least minimise it. Blinding, whereby the identity of the treatment assigned is concealed from patient and investigator, e.g. by use of a placebo designed to be indistinguishable from the experimental treatment, can be used to prevent subjective assessment bias, however this is often difficult to apply in cancer trials. The use of independent assessors blinded to the treatment arm of the patient can help to minimise subjective assessment bias.

Stratification

Stratification is the deliberate balancing of known prognostic factors between the arms. Pre-stratification achieves this at treatment assignment; post-stratification achieves this at analysis whereby the method of analysis *adjusts for* or *eliminates the effect of* given factors.

If stratification is used to balance *known* prognostic factors then randomisation eliminates bias due to *unknown* prognostic factors.

In addition, post-stratification can 'improve' the efficiency of the analysis. If a post-stratification factor is at least moderately strong the power of the comparison of the arms can be increased. Thus if pre-stratification factors are used consideration should be given to post-stratify by these factors in the analysis.

Choice of stratification factors

There is no universal agreement as to whether and how stratification should be used at treatment allocation. A reasonable approach is to stratify by known strong prognostic

factors and any, such as institution, which would be desirable, but difficult, to adjust for in the analysis. If conventional stratification (balanced blocks) is used it is important that not too many factors be used if the situation whereby some strata contain few patients can result. This restriction does not apply to the minimisation method.

Minimisation is a method of assigning treatments to patients so as to balance the distribution of treatment arms across each level of each of several factors. Treatment assignment is necessarily performed at the time of randomisation.

Generalisability

The aim of the trial is to be able to apply the results of the trial to future patients. The type of patients to whom the results will apply will be determined largely by the selection criteria (and also the profile of patients actually entered into the trial) – broad criteria allowing selection from a heterogeneous population may require a larger sample to demonstrate efficacy of the new agent and will be applicable to a wider population. Alternatively, efficacy may be easier to demonstrate in a more homogeneous population, i.e. using more selective criteria but the results may not be as broadly applicable to future patients.

3 Objectives and Endpoints

Objectives need to be clear and specific and to be categorised into a primary objective (sometimes more than one), secondary objectives and exploratory objectives.

The reason for preferring one primary objective and one primary endpoint has to do with the problem of multiple comparisons. If, when the trial data are analysed, many statistical tests are performed and, hence, many p-values are generated, the chance of producing a false positive result can be quite high and to protect against this risk, one objective is declared to be primary and others are given less importance; any result with $P < 0.05$ among secondary objectives is accorded less credibility or requires much stronger evidence (P -value much smaller than 0.05) before claiming it as a definitive result. It is important, therefore, to ensure that the endpoint for the primary objective is chosen wisely. More than one primary objective or primary endpoint may be chosen; an adjustment for multiple testing may then be appropriate. More credibility can be placed on primary and less on exploratory objectives.

The term *endpoint*, in general, refers to an outcome used for evaluating an objective of a trial. In a therapeutic trial, an endpoint is an outcome presumed to be indicative of treatment effect. It is common usage to allow *endpoint* to refer to events (response,

recurrence), continuous outcomes (survival duration), or summary measures of outcome or outcome differences (response rate, hazard ratio).

It is important that the three sections of the protocol, Objectives, Endpoints and Statistical Considerations, be consistent with one another; i.e., that each section uses the same endpoints, employs the same terminology for them, and maintains the same distinctions between endpoints which are primary and secondary.

Endpoint definitions

All endpoints given in the objectives must be defined in this section. It is important that the endpoints are clearly defined and that, where applicable, standard definitions are adhered to (e.g. the RECIST guidelines for response in solid tumours).

The types and frequencies of the investigations required to assess the endpoints should be clearly specified in the protocol; it is important that the frequency and rigour of investigation of patients is the same for all treatment arms, especially where endpoints are subjective or difficult to detect.

Statistical Considerations

This section should explain how the endpoints will be used in the statistical analysis of trial data to evaluate the primary and secondary objectives of the trial.

The sample size for the trial is based on the primary endpoint. If there are two primary endpoints, the sample size will be the larger of the sample sizes corresponding to the two endpoints. It may be appropriate to include an assessment of the adequacy of this sample size for evaluating some of the secondary objectives.

In written reports of trial data, the results and their interpretation should normally reflect the primary, secondary and exploratory endpoints as defined in the protocol in order to avoid the charge of selective reporting of results.

Endpoint Definitions in Radiation Oncology

Broad categories of endpoints, with examples, are:

- Initial effect of treatment on disease: response rate.
- Long term effect of treatment on disease: recurrence, death.
- Side effects of treatment and disease: early toxicity, late toxicity, quality of life.
- Other outcomes: costs.

Common treatment effectiveness endpoints

Events indicative of treatment effectiveness commonly include complete response (CR) or persistent disease, recurrence, failure and death (perhaps cause-related).

These may be 'site-specific', e.g. local, nodal, distant, or combinations of these. The term *recurrence*, and perhaps *relapse*, should be reserved for trials in which all patients are rendered clinically disease-free by the initial treatment; or, at least, the terms should be reserved for patients achieving a complete response. *Failure* is often used generically to cover any of a number of ways in which the treatment could be said to have *failed*, and its definition may vary, accordingly, depending upon the site and stage of the cancer being investigated.

Similarly, definitions of *time-to-event* endpoints need also to be carefully defined; this means specifying the date from which the time is measured (start date), what constitutes an event (or events), what the time to the event is censored by, and to what group of patients the definition applies (usually this is 'all patients' but in some cases, e.g. time to late toxicity, may be a subset of patients). (See table below.)

There is often no consistency of definitions in reports of trials; e.g. *disease-free survival* is sometimes used to mean *time to progression* and sometimes *time to progression or death*. It is recommended that any term which includes 'survival' implies that death is one of the events.

The following table gives definitions for commonly used time-to-event endpoints. This format is often a useful one for providing the definitions in the protocol. Note that in the following it is assumed that a *close-out date* will be employed; this is a chosen date, usually the earliest of the dates of last follow-up of patients alive and not lost to follow-up, such that all follow-up beyond the close-out date is ignored. This is done to minimise bias arising from the possible earlier reporting of follow-up for patients who experience an event.

Time-to-event outcome	Events	From	For	Censored by ¹
Overall survival	Death	Randomization ²	All patients	Nil
Cancer-specific survival	Death from the cancer	Randomization ²	All patients	Death from non-cancer causes ⁴
Failure-free survival	Recurrence Persistent disease in any site Death	Randomization ²	All patients	Nil
Time to failure	Recurrence Persistent disease	Randomization ²	All patients	Death without prior failure ⁴
Time to local failure	Local recurrence Persistent disease	Randomization ²	All patients	Death ⁴ , distant failure ⁵
Duration of CR	Local recurrence	Date of CR	Patients with CR	Death ⁴ , distant failure ⁵
Time to late toxicity	Late toxicity	Treatment start ³	All patients receiving 'adequate' treatment	Death ⁴

¹ as well as the close-out date (or date of loss to follow-up).

² phase II trials may measure times from registration or treatment start.

³ alternatively, time to late toxicity can be measured from 90 days following the start of radiotherapy.

⁴ note that it is assumed that death as a censoring event is independent of the event of interest.

⁵ distant failure may or may not be defined as a censoring event for local failure (see below).

In randomised trials, times are measured from the date of randomisation, rather than the date of treatment start, because this is when patients on the treatment arms are truly comparable (i.e. when no systematic bias exists).

In general, the use of overall survival is preferable to that of cancer-specific survival because for the latter (i) it is often difficult to determine the cause of death, (ii) definitions vary as to what types of deaths to include as cancer-related, and (iii) the analysis of cancer-specific survival involves the possibly dubious assumption that cancer-related and non-cancer-related deaths result from independent processes. If cancer-specific survival is used, however, it is recommended that cancer 'events' include treatment-related deaths and deaths of unknown cause, in addition to deaths resulting from the cancer; death with, but not resulting from, the cancer would be a censoring event.

In many instances, estimates of event rates (e.g. local failure rates) or event-free rates using Kaplan-Meier curves are inappropriate. These instances arise when the censoring mechanism is not independent of the risk of the event of interest, which may be the case whenever the censoring event is other than the close-out date. The use of cumulative incidence rates arising from the method of competing risks has been proposed as an appropriate way of representing the risk of the event in the presence of competing events. As an example, risk of local failure should often be assessed using a competing risks approach. The events that compete with it are often distant failure and death without preceding failure. (Alternatively, distant failure may be ignored and only death be considered as a competing event.) When comparing cumulative incidence rates of local failure between treatment arms, it is important also to consider possible differences in cumulative incidence rates of distant failure and death from other causes.

Toxicity endpoints

Early and late toxicity should be defined using the current version of the CTCAE grading systems. Note that the analysis of rates of late toxicity strictly should take into account the length of follow-up on each patient; this means using an actuarial approach. However, little bias would be involved in the comparison of arms because follow-up would be very similar for both arms. However the use of crude rates (proportions) may make it difficult to compare estimated rates of late toxicity rates with other studies.

4 Sample Size Estimation in a Phase III Trial

For any clinical trial it is important to determine the number of patients that are required to participate in the trial: sufficient to enable the trial objectives to be satisfactorily addressed but not so excessive that patients are treated on ineffective trial treatments unnecessarily.

So, while estimation of the sample size required for a clinical trial is the role primarily of the trial statistician, nevertheless the clinical investigators have a significant role at several stages of this process. Sample size estimation involves selecting input parameters including estimates of likely outcomes of treatments and of appropriate differences in outcomes between treatments that are appropriate to try to detect - such considerations are mainly the responsibility of the clinical investigator. Ideally, the clinical investigator should also understand the other assumptions made by the statistician and agree that these are reasonable.

4.1 Phase III superiority trial with a survival-type endpoint

In the following the common fixed sample size, two-arm parallel trial with survival as the main endpoint. is being considered. The elements of sample size calculation will be described and formulae and tables to derive approximate sample sizes will be described. The aim is to provide appreciation of what is involved – the assumptions made and the processes in the calculation and the ability to produce approximate sample size estimates – rather than provide a primer on sample size calculation. An experienced statistician should always be engaged to provide the definitive sample size calculation (and associated considerations) for a trial.

4.1.1 The elements of the sample size calculation.

1. Set fixed parameters.

There are, what may be called the fixed inputs or parameters for the sample size calculation. These are the variables that are preset and the same for most trials, namely:

- The type I error rate (α ; also known as false positive error rate). This is the chance that we will allow for a significant result to be obtained (i.e. for the P -value to be significantly low – rejection of the null hypothesis) when the null hypothesis is true. This is almost always set to 5% ($\alpha = 0.05$), which implies that we will claim to have demonstrated that the new treatment is superior to the standard if $P < \alpha = 0.05$.

- The proportion of patients (p) in the standard treatment arm. This is nearly always 0.5 ($p = 0.5$).
- Whether the comparison will be a 2-sided or a 1-sided test. This is nearly always 2-sided, which means that we want to be able to conclude that the new treatment is significantly inferior as well as superior, depending on in which direction a significant difference lies.

There are now two steps to follow. (1) calculate the number of deaths (more generally, events) to be observed and then (2) calculate the number of patients to be accrued to the trial.

2. Calculate the number of events

For this calculation there are three inputs, or parameters, that will vary from trial to trial. These are:

- The difference, δ , between treatment arms that the trial is designed to detect, if present.
- The power of the trial ($1-\beta$). This is the chance that if δ is in fact the true difference the new treatment (N) will demonstrated to be superior to the standard (S) (i.e. when $P < 0.05$). The power is usually set to 80% ($1-\beta = 0.80$), although 90% power is not uncommon and, perhaps, should more commonly be used.
- The total number of deaths (D) requiring to be observed at the time of analysis.

The *difference*, in survival studies, is commonly specified in terms of the survival rates of the standard and new treatment arms at a given time. For the purposes of sample size calculation, however, these rates need to be converted to a log hazard ratio (see later) using the (natural) log. If R_S and R_N are the survival rates at a given time (it doesn't matter at what time), the hazard ratio, δ , is given by,

$$\delta = \frac{\log(R_N)}{\log(R_S)}$$

So if the trial is required to detect a hazard ratio (δ) corresponding to survival rates at 5 years of 50% for the standard arm and 60% for the new arm, this hazard ratio is $\delta = 0.737$. (Note that it doesn't matter whether logs are to the base 10 or to some other base.)

When any two of difference, power and number of deaths are set, the third can be determined via formula. Conventionally, one sets the power and the difference to be detected and calculates the number of deaths required. The formula is:

$$\sqrt{Dp(1-p)} \log(\delta) = Z_{1-\alpha/2} + Z_{1-\beta}$$

where $\log(\delta)$ is the natural log of the hazard ratio, δ , and Z_p is the abscissa corresponding to an area, p , under the normal (bell-shaped) curve to its left. (For the purposes of this calculation the sign of $\log(\delta)$ can be ignored.)

For $\alpha = 0.5$, $Z_{1-\alpha/2} = 1.960$ and for $1-\beta = 0.80$ and 0.90 , $Z_{1-\beta} = 0.842$ and 1.282 , respectively, giving,

$Z_{1-\alpha/2} + Z_{1-\beta}$	$1-\beta = 0.80$	$1-\beta = 0.90$
$\alpha = 0.05$	2.802	3.242

So, for a 1:1 allocation ratio ($p = 0.5$), $\alpha = 0.5$ and $1-\beta = 0.80$,

$$\sqrt{D} = \frac{5.603}{\log(\delta)}$$

or,

$$D = \frac{31.40}{\log^2(\delta)}$$

or, approximately,

$$D = \frac{32}{\log^2(\delta)} = 344$$

So for the example above, comparing 50% versus 60% survival rates at 5 years, the approximate number of deaths required to be observed is 344. (More accurately, $D = 337$, but the difference is small and errs on the conservative side.)

3. Calculation of the number of patients required

Finally, the required number of patients can now be calculated after the following parameters are given:

- The accrual rate to the trial (r , assumed constant)
- The length of follow-up once accrual has ceased (f)
- The form (shape) of the survival curve for the standard treatment.

The form of the survival curve is commonly taken to be exponential but this should be thought about carefully. Especially important is whether the survival curve plateaus. (This more likely when dealing with a time to recurrence endpoint.) A survival curve

which exhibits better survival will require more patients to be accrued in order to provide the required number of deaths. On the other hand, a slower accrual rate, while it will increase the length of the trial, will require fewer patients because each patient will be followed, on average, for longer period of time and, hence, will be more likely to die within the duration of the trial.

The calculation of the required number of patients is not straightforward and usually requires a computer program to work it out. An approximate estimate can be obtained by working out the approximate probability of a trial patient dying by time t = the average potential follow-up time of a patient in the trial = half the accrual period + the follow-up period, by using the following formula:

No. of deaths observed = (No. of patients accrued) × (Proportion of patients dying)

$$D = ra \{1 - S^{(f+a/2)/t}\}$$

where r is the accrual rate, a is the accrual duration, S is the average survival ($\frac{1}{2}(S_0 + S_1)$) of all patient at time t and f is the follow-up duration from end of accrual. (Note that $f+a/2$ is the average potential follow-up time of all patients.) To obtain the accrual duration, a , and hence the number of patients, find, by trial and error in the equation above, the value of a that gives the required number of deaths.

If $r = 150$, $f = 5$, and, from above, $S = 0.55$ and $t = 5$, then $a = 4.0$. (This gives $D = 340$, close to the required 337.) The required number of patients is therefore approximately 600.

Cautionary note:

A sample size calculation to be included in a trial protocol should be the responsibility of an experienced statistician. The above approximate formulae are usually sufficiently satisfactory to provide a preliminary estimate of the required patient numbers for a trial but should not be trusted as a definitive estimate. Trial features that demand care in the sample size calculation are:

- Survival curve shape that is not exponential, especially one that plateaus
- Non-proportional hazards (the HR is not constant at all timepoints along the control and experimental curves)
- Competing events (e.g. when the endpoint is, say, time to local relapse)
- The event rate is low (e.g. comparing 90% versus 95%) local relapse-free rates at 10 years.

In fact, because most sample size programs use formulae that are ‘asymptotic’ and not exact, any substantial deviation from the normal case should be handled with care. The ‘gold standard’ method is simulation.

4.1.2 Sample size checklist

Quantity	Symbol	Example
Fixed parameters		
Type I error (false positive rate)	α	0.05
Proportion in control arm	p	0.5
One-sided or two-sided testing	s	2
No. of events required		
Control arm survival rate at a given time	S_0	50% at 5 years
Experimental arm survival rate at a given time	S_1	60% at 5 years
<i>Calculate hazard ratio</i>	HR	0.737
Power	$1-\beta$	80%
<i>Calculate no. of events (deaths)</i>	D	337
No. of patients required		
Shape of control arm survival curve		Exponential
Accrual rate (patients per year)	r	150
Duration of follow-up following end of accrual (years)	f	5
<i>Calculate duration of accrual period (years)</i>	a	4.0
<i>Calculate no. of patients</i>	N	600

Simulation suggests a figure of 610 is the required number. However, it should be realised that any sample size calculation is only approximate anyway, because:

- The main determinant of the required power is the no. of events (e.g. deaths) rather than the number of patients; this number can vary according to (i) chance, (ii) the accrual rate (there are more events, on average, if accrual is slower, (iii) the true survival rates (more events if lower than the postulated survival rates).
- There are usually losses to follow-up and patients whose management doesn’t comply with the protocol specifications (partially evaluable patients).
- Input parameters may be inaccurate.

The implications of these features are that:

- It is desirable, therefore, to inflate the sample size, say by 5% to 10%, to compensate for these influences.

- A target sample size should be rounded up, in any case, to at most two significant figures ($N = 453$ represents false accuracy).
- It is statistically preferable to cease follow-up when the required no. of events are reached rather than after a fixed follow-up time.

4.1.3 Sample size adjustments

Relative to the sample size required for a two-arm trial with 1:1 allocation, type I error of 5% and power of 80%, the following adjustments apply for variation in these conditions:

For	Change to required D
90% power	Increase by 34%
One-sided test	Reduce by 21%
Proportion, p , of patients in the standard arm	Increase by a factor, $1/4 p(1 - p)$

If, for example, a 1:2 allocation ratio was implemented the no. of events would need to be increased by a factor of 12.5%, compared with 1:1 allocation.

4.1.4 Sample size and a response rate outcome

This section applies to any outcome which is expressed as a proportion, such as an adverse event rate. The following table gives the total sample size required to detect a difference for two given response rates with power of either 80% or 90%.

Table 2. Total Sample Size for Comparing Two Proportions (e.g. Response Rates)

The entries in the tables give the total sample size required in order to compare two proportions (e.g. two response rates), P_1 and P_2 (where $P_2 > P_1$), from two equal-sized groups with a given power (either 80% or 90%), when using a two-sided test and type I error rate of 0.05. Proportions and their differences are given as percentages.

Smaller Proportion (%)	Difference between proportions (%)										
	5	7.5	10	12.5	15	17.5	20	22.5	25	27.5	30
5	850	430	270	190	140	110	90	74	64	56	48
10	1400	660	390	270	200	150	120	98	82	70	60
15	1900	850	500	330	240	180	150	120	96	82	70
20	2200	1100	590	390	280	210	170	130	110	90	76
25	2500	1200	660	430	310	230	180	140	120	96	82
30	2800	1300	720	470	330	250	190	150	120	100	84
35	3000	1400	760	490	340	250	200	160	130	110	86
40	3100	1400	780	500	350	260	200	160	130	110	84
45	3200	1400	790	500	350	260	200	150	120	98	82
50	3200	1400	780	500	340	250	190	150	120	94	76
55	3100	1400	760	480	330	240	180	140	110	86	70
60	3000	1300	720	450	310	220	170	130	96	76	60
65	2800	1200	660	410	280	200	150	110	82	64	48
70	2500	1100	590	360	240	170	120	86	64	46	
75	2200	930	500	300	200	130	90	60			
80	1900	760	390	230	140	86					
85	1400	550	270	140							
90	850	300									

80% power

Smaller Proportion (%)	Difference between proportions (%)										
	5	7.5	10	12.5	15	17.5	20	22.5	25	27.5	30
5	1200	570	360	250	190	150	120	100	84	74	64
10	1900	880	530	360	260	200	160	130	110	94	80
15	2500	1200	670	450	320	250	190	160	130	110	94
20	3000	1400	790	520	370	280	220	180	150	120	110
25	3400	1600	880	580	410	310	240	190	160	130	110
30	3700	1700	960	620	440	330	250	200	170	140	120
35	4000	1800	1100	650	460	340	260	210	170	140	120
40	4200	1900	1100	670	470	340	260	210	170	140	120
45	4200	1900	1100	670	470	340	260	210	170	140	110
50	4200	1900	1100	660	460	330	250	200	160	130	110
55	4200	1900	1100	640	440	320	240	190	150	120	94
60	4000	1800	960	600	410	290	220	170	130	110	80
65	3700	1700	880	550	370	260	190	150	110	84	64
70	3400	1500	790	480	320	230	160	120	84	60	
75	3000	1300	670	400	260	180	120	80			
80	2500	1100	530	310	190	120					
85	1900	730	360	190							
90	1200	400									

90% power

(Sample size calculation is based on the arcsin-root transformation.)

4.1.5 Time to recurrence endpoints and competing events in sample size calculation

When the endpoint of interest, say local recurrence, is subject to competing events (e.g. distant recurrence) – i.e. local recurrence is counted only when it occurs as (or as a component of) a first recurrence and any local recurrence occurring after a distant recurrence is ignored – the rate of censoring by the competing risk must be allowed for in the calculation of the sample size. This means that the number of events observed will be fewer than would occur without censoring by the competing risk and hence the number of required patients will be greater. Special software is needed to perform the sample size calculations.

4.2 Phase III trials – Non-inferiority trial

In a non-inferiority trial the aim is to demonstrate that the new treatment is no worse than the standard arm, with respect to the main endpoint. It is not possible to demonstrate that two arms are exactly equivalent so the formulation of the problem

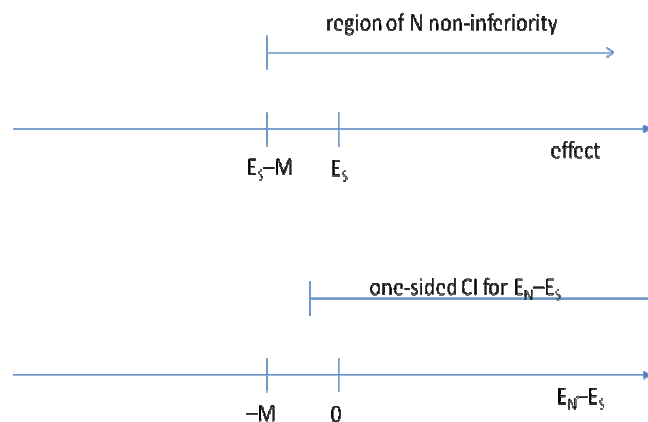
involves specifying a margin M such that if E_S is the effect of the standard treatment and E_N is the effect of the new treatment, the non-inferiority is said to have been demonstrated if it can be shown that $E_N > E_S - M$.

M is called the non-inferiority margin (N-IM) and is defined as the maximum 'difference' between the arms consistent with the assertion that the new treatment is *effectively* non-inferior to the standard treatment.

Two approaches are common and are equivalent: $E_N - E_S$ non-inferior

- (i) using a confidence interval for the difference between the arms, and
- (ii) using hypothesis testing.

A one-sided 95% confidence interval (CI), say, for the difference (New - Standard) of the form (0 to upper-limit), (or, equivalently, a two-sided 90% CI for the difference) is calculated and if this CI lies entirely above $-M$ it will be concluded that the new treatment is, effectively, non-inferior to standard treatment.



The equivalent formulation in terms of hypothesis testing is to test the null hypothesis,

$$H_0: E_N - E_S \leq -M \text{ (inferiority),}$$

versus the one-sided alternative hypothesis,

$$H_1: E_N - E_S > -M \text{ (non-inferiority),}$$

For a survival endpoint it is convenient to specify effects and margins in terms of survival rates and their differences. However, for the purposes of calculation of sample size and analysis of the data, a hazard ratio (standard:new) corresponding to these postulated survival rates will be used. A $HR = 1$ means exact equivalence and a $HR < 1$ means the new treatment is inferior. The region of non-inferiority will be those hazard ratios greater than the N-IM HR which will be less than 1.

In a non-inferiority trial it is most important to ensure that data quality is very high: any noise in the data would have the effect of narrowing any difference between the

groups being compared, thereby increasing the likelihood of demonstrating equivalence or non-inferiority when it does not exist - i.e. increasing the chance of a false positive result (of non-inferiority).

Hence, analysis should include both (i) an ITT analysis and (ii) confirmatory analyses using per-protocol patient subsets. A conclusion of non-inferiority would be made if both types of analyses were consistent in indicating non-inferiority.

4.3 Phase III trials sample size for a continuous outcome variable

The example of a QoL endpoint will be used.

Similar to a survival outcome:

- set fixed parameters (α , p , 1- or 2-sided)
- set variable parameters – difference (δ), within-group SD (σ), power ($1-\beta$), number of patients N . Choose two of: the effect size (δ/σ), power and N and determine the third using the following formula:

$$\frac{1}{2}\sqrt{N} \frac{\delta}{\sigma} = z_{1-\alpha/2} + z_{1-\beta}$$

In the example, suppose it is required to design a trial which is to have 80% power ($1-\beta = 0.80$) to detect a difference of 6 units (δ) in the scale of a QoL variable. Suppose also that the standard deviation of the QoL variable within each treatment arm is 15 units. The effect size (δ/σ) is 0.4 (6/15). Set the false positive rate: $\alpha = 0.05$. The sample size required is, by the above formula, $N = 196$.

An approximate formula for the total sample size (N) with a continuous variable, for which power is 80%, there is 1:1 allocation, $\alpha = 0.05$, and 2-sided testing, is:

$$N = \frac{32}{(\delta/\sigma)^2}$$

which gives $N = 200$.

5 Analysis and Interpretation of Trial Data

5.1 Interim analyses and premature termination

Interim reports should be prepared annually for all trials, for presentation at TROG Annual Meetings. These reports inform the trial group of progress on the trial, in order to promote interest and so encourage accrual, and monitor accrual to the trial.

Interim reports to the group should not include a comparison between treatment arms of the important trial outcomes (e.g. response rate, survival), even in blinded form where the treatments are not identified. Reporting toxicity rates by arm should not be made if this could have an effect on trial progress. The principle should be that an analysis should not be publicly presented which may haphazardly affect trial progress and outcomes (e.g. decrease accrual, change the profile of the patients entered, initiate significant protocol changes) – review of such analyses should be the preserve of the Data Monitoring Committee (DMC).

In the special case of group sequential trials, however, planned interim analyses are performed comparing the treatment arms with respect to the main outcome measure. However these analyses are not reported to the Group, nor even to the Trial Management Committee, but rather to the independent Data Monitoring Committee established for the purpose of administering the trial's stopping rule.

5.2 Group Sequential trials and Stopping Rules

It is important for investigators conducting clinical trials, particularly phase III trials, to include in their protocols a schedule of interim analyses. These planned interim analyses are usually done to either detect a significant difference in efficacy between the treatment arms as soon as possible in order to avoid prolonging use of an ineffective treatment, or to assess whether the toxic effects from one or more of the treatment arms are excessive.

Conducting analyses using a 5% significance level at each of the interim analyses can cause problems by increasing the chance of producing a false positive result. If you reject a null hypothesis at the 5% level, there is a 5% chance that you are wrong when the null hypothesis is actually true. If you test two independent null hypotheses and reject either one at the 5% level, the chance of getting a false positive result is nearly 10%. With successive analyses of accumulating data, the chance of a false positive increase but not linearly, because the results from one analysis will be correlated with the results from earlier analyses. Nevertheless the chance of drawing the wrong conclusion and stopping the trial unnecessarily early can become quite high.

This can be illustrated in the table which shows that if you conduct four analyses using a 5% significance criterion for stopping the trial at any analysis, the chance that

you will stop the trial and declare one arm superior when in fact no true difference exists is 13%.

No. of repeated tests Percentage significant at 5% level

1	5%
2	8%
3	11%
4	13%
5	14%
10	19%

One way of reducing the chance of reaching a false positive conclusion and still maintaining the overall significance criterion for the trial at 0.05 is to require the significance criterion for each interim analysis to be much lower than 0.05. There are a number of methods used for specifying *P*-values for stopping a trial at interim analyses so as to preserve the overall type I error rate at 5%.

The O'Brien and Fleming method is commonly used. This method is illustrated with an example. If you were planning a trial comparing response rates between two treatment arms with a total sample size of 400 patients and you decided to carry out three interim analyses after accruing 100, 200 and 300 patients respectively, then to achieve an overall significance level of 0.05, your criteria for rejecting the null hypothesis and stopping the trial at any one of the four analyses could be as follows:

Analysis P-value criterion

1	<0.00004
2	<0.0039
3	<0.018
4 (final)	<0.041

Using such a strategy maintains the overall false positive rate for the trial at 5%. A number of choices are available for selecting the interim *P*-values.

The O'Brien-Fleming rule was incorporated into the TROG neuropathic pain trial where there were two major interim analyses planned. The stopping rule was $P <$

0.0006 for the first analysis of 90 patients and $P < 0.015$ for the second analysis of 180 patients, leaving $P < 0.047$ as the significance criterion to be applied in the final analysis of 270 patients.

In deciding whether to stop a trial prematurely, all relevant information should be taken into account. In particular, strict observance of the statistical stopping rule is usually not a sufficient condition for termination: other factors usually need to be taken into consideration. The data for the analysis would often not be of as high quality as would occur at final analysis; if the endpoint is a time-to-event variable, the curves may converge with longer follow-up; other comparisons, such as of toxicity or quality of life outcomes (which should often be performed in interim analyses), may suggest that there are trade-off of benefits between outcomes.

5.3 Main Analysis and Publication

In regard to the main analysis of a trial ensure that the analyses of the trial data:

- meet international standards of reporting.
- are performed according to the protocol (and Statistical Analysis Plan if such exists in a separate document) specifications
- are conducted according to the planned timeframe (the analysis timetable should appear in the protocol, or in a Statistical Analysis Plan document, if one exists).

Once accrual to the trial has finished the following steps should be undertaken.

1. Data validation and planning for closeout.
The closeout date for the trial should be determined by specifying either (i) the end of a pre-planned period of follow-up following close of accrual or (ii) the anticipated time of occurrence of the number of events required to provide the power of the trial. Ideally, patients still alive and being followed should be planned to have their final trial follow-up visit occur within a short period of time (say 3 to 6 months) following the closeout date.
2. Final data collection
3. Planning of the analysis
4. Performing the analysis, including writing an analysis report.
5. Writing the manuscript based on the analysis report

6 Miscellaneous topics

6.1 P-values and confidence intervals

P-value

The p-value comes from the testing of null hypothesis versus an alternative hypothesis. Typically, if δ represents the true difference between the new and the standard treatment (new – standard) then the null hypothesis is: $H_0: \delta = 0$ versus $H_1: \delta \neq 0$. The sample size for the trial is determined such that there is a reasonable chance (usually 80%) that a specific alternative difference, say $\delta = \delta_0$ can be detected.

The *P*-value is the probability that, given that the treatments are truly identical, the observed difference between the treatments, or one more extreme, could have arisen by chance.

Confidence interval

The definition of (say) a 95% confidence interval for a population quantity, such as the true difference between two treatments, is commonly defined as one which contains the quantity with probability 95%. (Note that this is not strictly true but is close enough to the truth. Strictly, the 95% level of confidence is associated with the *method* of calculating the interval – the method is such that, on average, 95% of all confidence intervals generated by it will cover the true difference.)

Relationship between P-value and confidence interval.

When testing for a difference between two treatment arms, $P < 0.05$ if and only if a 95% confidence interval for the true difference does not include zero.

6.2 Interpretation of negative trials

It is not correct, in general, to conclude that a statistically non-significant result means that the treatment arms are effectively the same. Equivalence is demonstrated by (a) defining, *a priori*, an equivalence margin, say Δ , and (b) seeing whether a confidence interval (CI) for the true difference lies within the interval $-\Delta$ to $+\Delta$. This equivalence region is that containing all differences between treatment arms considered to be of minimal clinical significance. The CI can be a 95% CI or a 90% CI.

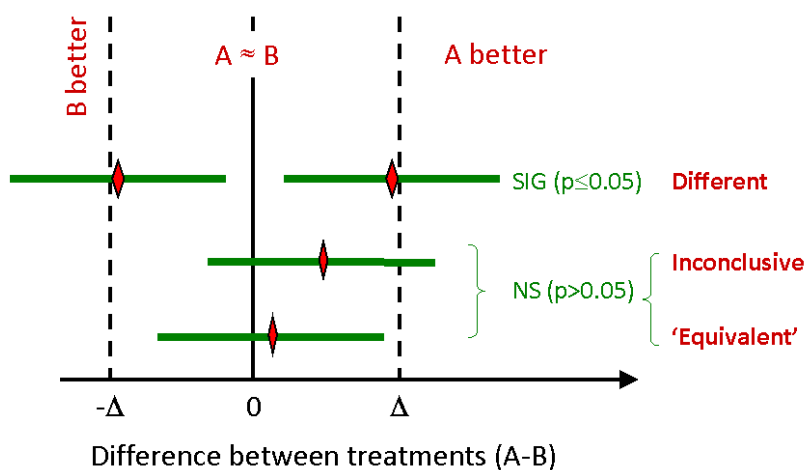
It is sometimes said that if an adequately-powered comparative trial does not reach statistical significance then it can be claimed that the two arms being compared are effectively equivalent. This is generally a wrong interpretation, for two reasons.

First, the target difference upon which the comparative trial was powered is not the same as the equivalence margin – nearly always it is somewhat greater. Second, the criterion for equivalence stated above (95% CI lies within the equivalence region) is not always satisfied. In fact, if the target difference, δ , were the same as the

equivalence margin, it can be shown that demonstrating equivalence requires that the $P > 0.40$ (if the power of the comparative trial is 80%) or $P > 0.20$ (if power is 90%). Only if the power of the comparative trial is 97.5% is it true that $P > 0.05$ implies equivalence (but only when target difference = equivalence margin).

It may be useful to define the equivalence margin *a priori* in the protocol in order to add credibility to any statement of equivalence in the trial publication.

Three possible outcomes (ideal case)



10

6.3 Small under-powered trials

The question of whether small underpowered trials are justifiable is controversial. One point of view is that if there is no ethical issue and patients are informed that the trial may not result in a definitive conclusion, an important consideration is that of resources: “does the investment of the resources required to run the trial justify the limited conclusions that can be made from the trial?” Such a trial should be registered (and hence be available for inclusion in a meta-analysis). All trials should be published, if possible, whether conclusive or not. From a scientific point of view, some information is better than no information. The danger of misinterpretation of negative result of a small trial should, perhaps, be regarded as an issue of education rather than misinformation.

6.4 The hazard ratio and the Cox proportional hazards model

The proportional hazards model (PHM) is commonly used as the basic model for analysis and interpretation of a randomised trial for which there is a survival-type endpoint. Such a model has an intuitive interpretation but is chosen mainly for its mathematical tractability. Departures from, proportional hazards, unless moderate to severe, do not matter a great deal to the validity of the analysis. The following is a simplified description of the PHM.

Imagine the true (population) survival curve of the control arm patients. Suppose the time axis is divided up into relatively small periods of time, say months, and consider the death rate for each month: the expected number of patients dying during the month as a proportion of those still alive at the start of the month. The death rate is a particular example of the more generic term, hazard rate – it could be a relapse rate in another example. Each month would have its own monthly death rate, usually quite small, e.g. 1% per month. As a simple example, suppose in the control arm every month had a 1% death rate, the survival rate at 5 years would be $55\% = 0.99 \times 0.99 \times 0.99 \times \dots \times 0.99 = 0.99^{60}$. That is, if there were 100 patients at the beginning, there would be 99 patients alive at the end of one month, then 99% of 99 patients alive at the end of two months, and so on.

Now imagine the death rates of patients in the experimental arm. If the PHM describes the relationship between the two survival curves, the death rates in corresponding months will be proportional to one another; i.e. the death rate in the experimental arm for a given month = $k \times$ the death rate in the control arm for that month. Suppose the experimental arm halves the death rate per month ($k = 0.5$), i.e. to 0.5% per month. The 5-year survival rate would be $0.995^{60} = 0.74$.

In this example $0.74 \sim 0.55^{0.5}$ and in general if the intervals were arbitrarily small (e.g. one day or even less) the survival rates would have this exact relationship:

$S_{\text{exp}} = S_{\text{std}}^{\text{HR}}$, where HR (k in the above example) is the hazard ratio (death rate ratio in this case).

Thus, the definition of the PHM is that ‘instantaneous hazard rates are proportional at all times points and this occurs if and only if the survival curves are related by

$S_{\text{exp}} = S_{\text{std}}^{\text{HR}}$, for some value HR, which is the hazard ratio.

It follows that if two survival curves, say, conform to the PHM and the survival rates at some given time are S_{std} and S_{exp} , the implied hazard ratio is given by $\text{HR} = \log(S_{\text{exp}}) / \log(S_{\text{std}})$.

6.5 Analysis of QoL data

Refer to the TROG Guideline TPP E7 Assessing Health Related QoL for statistical aspects of QoL evaluation. See also section 4.3 above.

6.6 Rules of Thumb

6.6.1 Confidence interval for a proportion when none (or all) patients respond

For 0 out of n responses, 95% CI is approximately 0 to $4/(n+4)$.

Ex.: 0 responses in 16 patients: 95% CI is 0 to 20%.

For n out of n responses, 95% CI is approximately $n/(n+4)$.

Ex.: 16 responses in 16 patients: 95% CI is 80% to 100%.

6.6.2 Approx sample size for comparing two binary proportions (e.g., response rates)

For a power of 80%, 1:1 allocation, $\alpha = 0.05$, the approximate size of each group =

$$n = (200/\%difference)^2.$$

Response rates must be around 50%, e.g. between 30% and 70%.

Ex. To detect a difference corresponding to response rates of 30% versus 40%, i.e. a difference of 10%, $n = (200/10)^2 = 400$ per group, or 800 patients in total.

6.6.3 Relationship of P-values to outcomes

When a trial has 80% power to detect a difference in a parameter θ (e.g. the difference in two response rates) and t is the estimate of θ from the trial data, then,

$$\text{If } t = \theta, P = 0.005$$

$$\text{If } P = 0.05, t = 0.7 \times \theta$$

In a small trial with power of only 50% to detect θ ,

$$\text{If } P = 0.05, t = \theta$$

Note that these are theoretical results and will be approximate in practice.

Ex. A trial was designed to have 80% power to detect a difference of 15% (θ) in response rates between two treatment arms. At the end of the trial, the main analysis of the data showed that 15% (t) was in fact the observed difference in response rates between the arms. This means that we would expect that the P -value for testing this difference would be about $P = 0.005$.

7 Acknowledgements

TROG gratefully acknowledges the ongoing commitment of TROG's Statistical Advisor, Assoc Prof Richard Fisher (Centre for Biostatistics and Clinical Trials, Peter MacCallum Cancer Centre) in the detailed review and revision of these guidelines.

8 References

1. Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995; 72: 511-518.
2. Ashby, D and Machin, D. Stopping rules, interim analyses and data monitoring committees. *Br J Cancer* 68:1047 - 1050, 1993.
3. Bailar III JC. Science, statistics and deception. *Ann Int Med* 1986; 104: 259-260
4. Gelman R, Gelber R, Henderson IC, Coleman CN, Harris JR. Improved methodology for analysing local and distant recurrence. *J Clin Oncol* 1990; 8 (3): 548-555.
5. Hait, WN. Updated Methods for Reporting Clinical Trials. *Clin Cancer Res* 2005;1 6753 1(19) October 1, 2005
6. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *Br Med J* 1996; 313: 36-39. Correction: 1996; 313: 550.
7. Keech Anthony, Gebiski Val, Pike Rhana (eds). *Interpreting and Reporting Clinical Trials: A guide to the CONSORT statement and the principles of randomised controlled trials*. Australian Medical Publishing Company, 2007.
8. Lewis JA and Machin D. Intention to treat – who should use ITT? *Br J Cancer* 1993; 68: 647-650.
9. McPherson, K. Statistics: the problems of examining accumulating data more than once. *New Engl J Med* 290:501 - 502, 1974.
10. O'Brien, PC and Fleming, TR. A multiple testing procedure for clinical trials. *Biometrics* 35:549 - 556, 1979.
11. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet* 2002; 359: 1686-9.
12. Pocock, SJ. When to stop a clinical trial. *Brit Med J* 305:235 - 240, 1992.
13. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials* 17: 343-346, 1996.

14. Simon, R. Optimal two stage designs for phase II clinical trials. *Controlled Clinical Trials* 10:1-10, 1989.
15. ICH Harmonized Tripartite Guideline: Statistical Principles for Clinical Trials (ICH E9). Available at:
http://www.tga.gov.au/docs/html/euguide/euad_clin.htm#clinicalgeneral
16. ICH Harmonized Tripartite Guideline: Structure and Content of Clinical Study Reports (ICH E3). Available at:
http://www.tga.gov.au/docs/html/euguide/euad_clin.htm#clinicalgeneral
17. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. *Ann of Int. Med.* 2001; 134(8): 663-694. (See also, <http://www.consort-statement.org>)
18. Altman DG, et al., for the CONSORT Group. The Revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Int Medicine* 2001; 134(8); 663-694.