Cornerstones of a well-designed phase III trial

Marc Buyse

IDDI (International Drug Development Institute), 430 Avenue Louise, B14, 1050 Brussels, Belgium

Introduction

Phase III clinical trials are considered the "gold standard" to demonstrate the effects of experimental therapies compared with standard therapy for the disease under consideration. For example, new drugs must typically be shown to have a sufficient level of efficacy and safety in two independent phase III trials before they are approved for marketing by the health authorities. Likewise, new treatments are rarely adopted in clinical practice if they have not been tested in phase III trial(s).

The purpose of this paper is to discuss the cornerstones of a well-designed phase III trial. The main sections of a typical phase III trial protocol are discussed in turn, and for each of them, key questions are identified that should be addressed when designing a phase III clinical trial, or when assessing its results.

Trial design

What is the hypothesis of interest?

The purpose of a randomised trial is to test a statistical hypothesis. Most common is the hope is to show that the experimental group is better than the control group in terms of an efficacy endpoint of primary interest (such as time to disease progression), even though there are inevitably other endpoints that are looked at (such as toxicities to treatment). The statistical approach to showing superiority of the experimental treatment is to test a *null hypothesis* of no difference, in the hope that the data collected in the trial will convincingly demonstrate this null hypothesis to be incompatible with the data, in which case the null hypothesis will be rejected. For instance, it would be unlikely for times to progression to be longer in the experimental group than in the control group if there was no real difference between the treatments. The *P-value* of the statistical test carried out at the end of the trial quantifies the probability of a difference as large as that observed if the null hypothesis were true. If the P-value is less than some pre-specified probability (referred to as the "level of significance" or "α-level"), then the result is said to be statistically significant. Thus, a P-value of less than 5% indicates that it is rather unlikely (less than 1 chance in 20) that the observed treatment difference is merely due to chance rather than to a true treatment effect, and this is conventionally considered to be statistically significant.

When the control group of a randomised trial is an active therapy considered standard, the aim of the trial may be to show that the efficacy of the experimental treatment is not inferior to that of the standard treatment, while being less toxic or more convenient. In this case, the null hypothesis is that the experimental treatment does *worse* than the control group, and again one hopes to be able to reject this null hypothesis. The ATAC trial, for instance, was designed to show non-inferiority of anastrozole compared to tamoxifen alone in terms of disease-free survival, but superiority of the combination of anastrozole and tamoxifen over tamoxifen alone. As it turns out, the trial showed that anastrozole was significantly superior to tamoxifen alone, while the combination was no better than tamoxifen alone. These were rather unexpected results compared to the pre-specified hypotheses, but because of the large number of patients included in the trial (over 9,000), these results were established with great statistical confidence [1].

How are patients randomised?

In phase III clinical trials, patients are allocated by a chance mechanism (*randomisation*) to receive one of the therapies being compared. The fundamental feature of randomisation is to provide comparability of the treatment groups with respect to all known and unknown factors, thus permitting an unbiased comparison between the treatment groups. Many benefits

follow directly from randomisation: simple, unadjusted statistical tests provide valid treatment comparisons that are more likely to be convincing than adjusted comparisons based on elaborate models; changes over time in the patient population under study, in diagnostic procedures, or even in evaluation of therapeutic response will affect all randomised groups equally, and will therefore not invalidate the treatment comparisons.

The way in which the various treatments are allocated to the successive patients who enter a phase III study must be carefully defined. Simple randomisation consists of choosing the treatment at random, regardless of patient characteristics. The advantage of simple randomisation, beside simplicity, is that it completely eliminates *selection bias* since the next treatment assignment is never predictable. The disadvantage of simple randomisation is that it does not protect against an *accidental bias* that may occur as a result of chance imbalances between the different treatment arms [2].

Stratified randomisation consists of allocating treatments after taking account of important patient characteristics, called *stratification factors* (e.g. age, gender, disease stage, prior therapies, etc.) The purpose of stratified randomisation is to reduce the likelihood of chance imbalances in the treatment assignments among strata. When stratified randomisation is used, the potential for accidental bias resulting from imbalances between treatment groups is reduced, the results of the trial may be more convincing because treatment groups look alike in terms of the important prognostic factors, and the precision of the estimates of treatment effect is increased (although the gain is usually quite small). If stratification is adopted, it is advisable to stratify only for factors of known prognostic value. If, for instance, histology were of no prognostic impact on the patient outcome, it would be pointless to stratify for histology. It is also advisable to stratify only for factors that are known with certainty at the time of randomisation. If, for instance, histology were only confirmed several weeks after randomisation, it would be hazardous to stratify for histology. In multicentre trials, it is also advisable to treat centre as a stratification factor to limit treatment imbalance within each centre.

Minimisation is a dynamic process which takes into account the distribution of prognostic factors of patients already randomised when allocating a treatment to a new patient, in order to minimise the risk of an imbalance between the treatment groups with respect to these prognostic factors. The major advantage of using minimisation is that good treatment balance can be achieved across a large number of stratification factors: for instance, in a multicentre trial comparing several anti-emetic therapies, six factors predictive of emesis as well as centre were taken into account using minimisation, and the distributions of these factors were almost identical among all treatment groups [3].

Treatments

What is the control group?

When designing a randomised phase III clinical trial, it is crucial to identify the appropriate control group to which the treatment group(s) will be compared. An untreated control group is indicated when no standard treatment exists for the disease under consideration, as may be the case in the adjuvant setting after resection of a solid tumour. Many early trials of adjuvant therapy compared an experimental treatment to no further treatment after surgery, but today, the control group typically consists of some standard treatment with established efficacy, called an *active control* group. Ideally, the control group should receive the standard therapy that would be given outside of a clinical trial setting, but such a standard does not always exist as practices may differ considerably across countries or even across hospitals within the same country. It is sometimes advantageous to let each hospital decide on the control group they feel most comfortable with, because this reflects actual clinical practice, rather than artificial trial conditions.

The most reliable form of control treatment consists of a *placebo*, which is seldom feasible in trials of cytotoxic agents, but should be considered with other agents such as cytokines, hormonal agents, etc. In such cases, the treatments may be given in *double-blind* fashion, whereby neither the physician nor the patient is aware of the treatment (otherwise, the treatment is said to be *open-label*). The main advantage of double-blind trials is that the assessment of endpoints is completely unbiased by knowledge of the treatment received. The recently reported ATAC trial, for instance, was a double-blind trial of tamoxifen alone, anastrozole alone, or the combination of both drugs. Thus, in this trial, every patient was randomised to one of three treatment groups: tamoxifen plus anastrozole placebo, tamoxifen placebo plus anastrozole, or tamoxifen plus anastrozole [1].

What are the experimental groups?

From a statistical standpoint, a simple design comparing two groups (an experimental and a control

group) is generally preferable to a design comparing multiple groups because the hypothesis of interest is simple, the interpretation of the trial results is straightforward, and the problem of *multiple comparisons* is avoided.

A notable exception to the simple design comparing two treatments is the *dose-ranging* design, in which patients are randomised between several doses of the same drug or combination of drugs. In this case, the aim of the trial is usually to test that the therapeutic response increases with dose. Statistically, this can be done through a test for the slope of the dose–response relationship, a powerful test that does not depend on the number of dose levels tested. When the goal of the trial is to find the most effective dose, however, each dose must be compared to control, and the problem of multiple comparisons must be addressed.

Another case where more than two treatment groups are desirable is *factorial designs*, in which patients are randomised more than once (although the different randomisations can occur concomitantly). Such designs are useful when two or more questions are simultaneously of interest for the same patient population. For instance, a recently reported trial simultaneously tested dose-dense (q 2 weekly) versus conventional (q 3 weekly) chemotherapy schedules, and sequential versus combination administration of the same agents. Thus each patient in this trial was randomised twice: first, between a dose dense or a conventional schedule, and second, between the sequential or the combination administration [4]. Under the assumption of no interaction between the two questions, a factorial design allowed the investigators to study these two questions with the same number of patients as they would have needed to study either question alone! In other words, studying each question separately would have required twice as many patients as studying them both in a single factorial design. That factorial designs should result in such huge savings in terms of patient numbers is somewhat counter-intuitive, but is merely due to the fact that every patient contributes to both questions independently. Sometimes, however, factorial designs fail because of an interaction between the two questions being investigated. For instance, a trial in colorectal cancer tested simultaneously 5FU + leucovorin versus 5FU + levamisole, and the duration of either regimen (6 versus 12 months). Unfortunately, the optimal duration seemed to depend on which of the two regimens (5FU + leucovorin or 5FU + levamisole) was administered, so that no general conclusion could be drawn from this trial! [5].

Yet another case where more than two treatment groups may be advantageous is when a new drug is being tested to either replace, or be added to, some existing standard drug. The ATAC trial provides a good case in point. A factorial design could not be considered for this trial because no patient could be left untreated, hence all patients received either tamoxifen or anastrozole, or both [1].

Patients

What is the target population?

The choice of the appropriate target population is often a matter of heated debate when designing a trial. Indeed, two conflicting arguments come into play: on the one hand, it makes sense to restrict the trial only to patients who may benefit from the intervention, while on the other hand, it seems sensible to open the trial to as many patients as possible, for in the absence of evidence to the contrary, all patients may *a priori* be assumed to benefit from the experimental treatment, albeit to varying degrees. We examine these two arguments in turn.

A "targeted" approach is warranted if it is known that only a subset of patients will benefit. For example, in patients with breast cancer, it is now well established that only patients expressing oestrogen (ER^+) or progesterone (PR^+) receptors benefit from hormonal therapy. It should be noted, however, that until recently patients not expressing oestrogen nor progesterone receptors (ER^- PR^-) were also assumed to benefit from tamoxifen through some cytotoxic (rather than hormonal) effect. The lack of benefit in hormone-receptor-negative patients has in fact been established reliably through inclusion of such patients in randomised trials. As tumour biology evolves and drugs are developed for specific genomic or proteomic targets, there will be more and more situations in which trials will be restricted to a patient population having a specific genetic profile. For instance, a trial in which trastuzumab (Herceptin) is used will be restricted to patients who express the *her-2-neu* gene. Thus the population of eligible patients will be reduced, but the treatment effect will likely be much larger than that obtained with non-targeted treatments such as anthracyclines or taxanes, and will therefore require less patients to be convincingly established. As of today, there are very few examples of targeted therapies that have indeed been shown to have a major impact on relevant clinical endpoints. However, these examples are encouraging enough to provide a model for future clinical trials. One celebrated example is that of imatinib mesylate (Gleevec, Glivec), a selective inhibitor of

the BCR-ABL tyrosine kinase used in chronic-phase myeloid leukaemia patients who are Philadelphia-chromosome-positive [6]. In these patients, imatinib mesylate produces a rate of major cytogenetic response of 87%, versus only 35% in patients treated with interferon-alpha plus cytarabine. At such levels of efficacy, large clinical trials are no longer needed to show small, incremental benefits.

In contrast to the targeted approach, a "broad" approach is warranted in the absence of definite knowledge about the factors predicting the therapeutic outcome. For cytotoxic drugs, for example, there is no biological reason to believe that the drug will work in some subsets of patients but not in others, and the decision to treat must therefore be based on the benefit/risk ratio for individual patients. In such cases, it is arguable that strict eligibility criteria are needed, for this may result in excluding patients who could benefit from the experimental treatment. A case in point is the arbitrary age limits that have too often excluded elderly patients from clinical trials, even if they were otherwise fit to receive either of the treatments under comparison. A better strategy is to let the participating physicians decide on which patients they enter in the trial, based on their clinical judgment. Statistically speaking, when there is doubt about which patients should be included, the choice between targeted or broad eligibility criteria can be based on considerations of sample size and trial duration. Let us take the example of adjuvant therapy for colorectal cancer. Assume a trial is being considered to compare the best available therapy to some experimental therapy. The trial will be open to all patients with Dukes' C tumours (with lymph node involvement), but the question is whether it should also be open to patients with Dukes' B tumours, under the plausible assumption that the relative treatment benefit is the same among patients with Dukes' B and with Dukes' C tumours [7]. Patients with Dukes' B tumours have a far better prognosis than patients with Dukes' C tumours, and therefore any treatment benefit is less easy to detect in these patients. This would argue against their inclusion. However, the trial will obviously take longer to accrue any given sample size if patients with Dukes' B tumours are excluded, and therefore there may be situations in which it is preferable to include them anyway. Moreover, it may be of interest to test the benefit of a new treatment in both Dukes' B and Dukes' C, even if it takes longer to show in the former than the latter. All in all, the only patients who should definitely be excluded from a clinical trial are those who are known not to benefit from therapy. At the present time, knowledge of factors that predict such lack of benefit is quite limited, but a better identification of molecular heterogeneity may soon have a substantial impact on clinical trial design and on the numbers of patients required [8].

How many patients are included?

The number of patients included in a comparative trial, called the *sample size* of the trial, must be sufficient to detect a difference deemed of clinical relevance. The sample size is calculated so as to guarantee that the difference of interest, if real, will be detected with a given probability, called the *statistical power* of the trial. In order to calculate a sample size, the trialists need to agree on the following design parameters: the significance level (α, usually taken equal to 5%), the statistical power (usually larger than 80%), the expected outcome in the control group and the desired outcome in the treated group (or a difference of interest between control and treated). Tables are available for different types of outcomes, and for different values of the design parameters [9].

Many trials in the past ended up being inconclusive (not showing a statistically significant difference between the treatment groups) because of an insufficient sample size (and, as a result, a low power). In this case, a meta-analysis of all related trials would be the best way of establishing real, but small, treatment differences [10]. Large-scale trials are justified and are now routinely carried out when the difference of interest is small. For instance, the ATAC trial randomised over 9,000 patients for the treatment of patients with early breast cancer. Such a large sample size was needed because the primary goal of the trial was to show that anastrozole was not inferior to tamoxifen in terms of disease-free survival (i.e. only a small difference between the two regimens would be accepted if it were against anastrozole), while being safer than tamoxifen in terms of drug-related endometrial cancer [1].

The sample size of a trial depends primarily on the difference of interest. This difference may vary greatly depending on the disease and the treatment considered. For instance, the trial comparing imatinib mesylate with interferon-alpha plus cytarabine in myeloid leukaemia was planned to detect a rather modest absolute difference of 10% in 5-year progression-free survival rates (assumed to be 50% in the interferon-alpha plus cytarabine arm versus 60% in the imatinib mesylate arm). In order to detect this difference, a sample size of over 1,000 patients was needed [6]. In the event, the benefit observed with imatinib mesylate far exceeded these expectations, since after only 18 months, the absolute difference in progression-free survival rates was already 18%

(74% in the interferon-alpha plus cytarabine arm versus 92% in the imatinib mesylate arm). In retrospect, such a huge treatment benefit could have been seen in far less than 1,000 patients, but the phase III trial had been planned conservatively to detect a smaller difference that would still have been of major clinical importance.

Are there subsets of interest?

A *prognostic factor* is a patient characteristic that modifies his or her prognosis: for instance, breast cancer patients with nodal involvement tend to fare less well than those without such involvement. A *predictive factor* is a patient characteristic that modifies the effect of a treatment: for instance, breast cancer patients without hormone receptors do not benefit from tamoxifen therapy, while patients with hormone receptor do. It is obviously of interest to identify subsets of patients who do not benefit from treatment, or conversely the subset that benefits the most, but the search for subsets is a perilous statistical exercise [11]. Indeed, in a clinical trial, the probability of finding a statistically significant result just by chance (if there were no real difference between the treatments being compared) is equal to α, the significance level. This level is often set conventionally at 5%, which means that on average one trial in 20 will falsely claim that a difference exists when there is none. This calculation assumes that just one comparison is performed. If multiple comparisons are performed, the probability is increased. Thus, if two subsets are looked at, three treatment comparisons are performed: one overall, plus one in each subset. If each of these comparisons is performed using the conventional 5% significance level, the overall significance level will be increased to more than 14%. If twenty subsets are looked at, the overall significance level will exceed 65%, and thus it will be more likely than not that at least one subset will show a "statistically significant" treatment difference that in fact does not exist. This explains why inappropriate subset claims create enormous confusion in the clinical literature.

The most reasonable way to interpret subset analyses is to examine the biological plausibility of the findings. Even when a plausible interpretation is put forth, it is desirable that the findings be reproduced in an independent series of patients before they are assumed true. For instance, some authors reported a striking association between the effect of chemotherapy, gender and tumour site in patients with Dukes' C colorectal cancer [12]. However, these results could not be reproduced in an independent series of patients receiving chemotherapy through a liver infusion (Table 1) [13].

In fact, based on the P-values of the tests within subsets, the results in the two series were exactly opposite: one series showed a statistically significant benefit in women ($P < 0.0001$) but not in men ($P = 0.13$), the other showed a significant benefit in men ($P = 0.04$), but not in women ($P = 0.21$), etc. It seems unlikely that these subset differences are real in any of these two series; it seems more likely that they merely result from the play of chance. In fact, *interaction tests* would have been a more reliable way of comparing subset results than reporting numerous subset P-values. Simon proposed useful guidelines to assess subset results (Table 2) [14].

Endpoints

How are the patients followed-up?

All patients who are randomised in a phase III trial should be followed up according to the study proto-

Table 1

Subset analyses showing discrepant results when performed in different groups of patients receiving adjuvant treatment after resection of a colorectal tumour ("significant results", i.e. $P \leq 0.05$, are in italics)

	Retrospective analysis of patients on chemotherapy [12]			Prospective randomised trials testing liver perfusion [13]		
Features	No. of patients	Hazard ratio	P-value	P-value	Hazard ratio	No. of patients
Right and left tumours						
Women	320	0.37	*0.0001*	0.21	0.80	293
Men	335	0.79	0.13	*0.04*	0.71	321
Right-sided tumours						
Women	146	0.30	*0.0001*	0.77	0.90	87
Men	113	0.40	*0.0007*	0.45	0.75	82
Left-sided tumours						
Women	174	0.45	*0.02*	0.14	0.73	206
Men	222	1.08	0.69	*0.05*	0.70	239

Table 2
Checklist to assess results from subset analyses (adapted from Simon, 1988 [14])

- Was subset analysis planned in protocol?
- Is subset analysis biologically plausible?
- Was subset analysis suggested by other prior evidence?
- Were subsets defined *a priori*? (Especially when a continuous variable defines the subsets, e.g. age <45 years vs. age ≥45 years)
- How many subsets were looked at?
- Are the subset results so unusually extreme as to rule out chance? (Note that this is best tested through an "interaction test")
- Was there any attempt to validate the results? (With other prospective series or even with historical data)
- Which statistical method was used?

col, even if they are found, after randomisation, to be ineligible or inevaluable for any reason. The most reliable analysis of a trial is based on the *intent-to-treat* principle, which includes all randomised patients, regardless of any protocol violations. In particular, patients who take other treatments have to be kept in the treatment group to which they were randomised. All other forms of analysis may be biased and, as such, are less desirable from a statistical viewpoint. In an intent-to-treat analysis, the number of patients who drop out of the trial prior to reaching the endpoint of primary interest should be kept to an absolute minimum.

A phase III trial protocol should be precise and detailed, but it should not attempt to provide exhaustive guidelines for patient management, since many of the routine examinations and procedures that would be performed outside of the clinical trial contribute no useful information to the endpoints of the trial. Likewise, in a phase III trial, it is generally undesirable to submit the patients to a more thorough or precise follow-up than what they would receive in routine clinical practice, so long as the endpoints of interest are assessed reliably.

Follow-up should be identical in thoroughness and frequency in the various treatment groups. For instance, seeing treated patients more frequently than control patients could bias the assessment of disease-free interval, because recurrences would be detected earlier in the treated group. Softer endpoints, such as disease recurrence, are more subject to bias than harder endpoints, such as death. For instance, if an untreated control group is compared to a treatment group, there may be pressure to scrutinise the untreated patients much more thoroughly than the treated ones in order to identify and treat disease recurrences as early as possible. When endpoints are subjective, they should ideally be assessed blindly, i.e. by investigators not aware of the treatment actually received, but this practice has limited applicability in clinical trials of treatments with noticeable side-effects and toxicities.

What are the endpoints of interest?

The ideal endpoint for a phase III trial is one that is important to the patient, observed soon after treatment inception, clinically meaningful, statistically sensitive to treatment effects, measured objectively and without bias. If such an endpoint existed, it could always serve as the *primary endpoint* of randomised trials (the primary endpoint is that used to calculate the sample size, and to determine whether the trial shows a significant effect of treatment or not). Unfortunately, no single endpoint fulfils all the desirable conditions. This is illustrated by some endpoints commonly used in advanced disease (Table 3).

Each endpoint, if chosen as the primary endpoint of a randomised trial, would have advantages and drawbacks: response to treatment (tumour shrinkage) is not sufficient, in and of itself, to establish patient benefit, time to disease progression is hard to measure objectively, and survival is insensitive to true treatment differences (Table 4) [15].

Usually, therefore, all of these endpoints are analysed and the totality of the evidence is taken into account to support claims of treatment benefit. Whenever possible, attempts are also made to measure the patient's quality of life, or at least some aspects of symptom-related quality of life. In some advanced forms of cancer (e.g. pancreas), the "clinical benefit"

Table 3
Essential features of endpoints used to assess therapies for advanced solid tumours

Endpoint	Clinical features:		Statistical features:	
	Time of occurrence	Relevance	Reliability	Sensitivity
Tumour response	Early	Low	Low [a]	High
Time to progression	Intermediate	High	Low [a]	High
Overall survival	Late	High	High	Low

[a] High if reviewed by independent panel blinded to treatment.

Table 4
Pros and cons of different endpoints used to assess therapies for advanced solid tumours

Endpoint	Pros	Cons
Tumour response	• Measured early (weeks to months) • Measured easily • Reflects biological activity • Assessment can be reviewed blindly by expert committee	• Responses infrequent • Insensitive to disease stabilisations (cytostatics) • Assessment prone to error and/or bias if not reviewed • Disease not always measurable • Limited impact on survival
Time to progression	• Reflects control of disease process • Unaffected by competing risks of death • Very sensitive to differences in treatment efficacy • Possible impact on survival • Closely related to quality of life	• Assessment subjective • Assessment potentially biased to allow for change in therapy • Assessment can be reviewed only after changes in therapy
Survival	• Most meaningful • Most objective	• Difficult to modify, therefore large sample sizes needed • Measured late (months to years) • Influenced by second-line treatments • Influenced by competing risks • Insensitive to short-term benefits

has been quantified using scales that combine performance status, weight loss and use of analgesics. Changes on such clinical benefit scales constitute meaningful outcomes to the patients and may be quite sensitive to real treatment effects. As such, they seem quite useful and often more relevant than general-purpose quality of life questionnaires.

Another issue of major interest is the identification of *surrogate* endpoints or markers, which, if valid, would allow trialists to replace a distant endpoint (such as the patient's death) by endpoints or markers that are observed earlier in the course of the disease (such as tumour shrinkage or sustained decreases in a tumour-related marker). For a surrogate endpoint to be valid, two conditions should be fulfilled: first, the surrogate endpoint must be predictive of the true endpoint for individual patients, and second, the treatment effect on the surrogate endpoint must be predictive of the treatment effect on the true endpoint for groups of patients (e.g. in successive trials). Unfortunately, few endpoints or markers qualify as valid surrogates for the endpoints of ultimate interest. For instance, in advanced colorectal cancer, tumour response is highly predictive of longer survival in individual patients, but the effects of treatment on tumour response and on survival are only mildly correlated [16]. Hence, even if an experimental treatment induced higher response rates in advanced colorectal cancer, its effect on survival would remain elusive. The discovery of markers that reflect relevant biological mechanisms or of the desired therapeutic effect may make the search for surrogate markers more promising in the future [17].

Some phase III trials address specific issues related to the cost effectiveness of the treatments under investigation, or translational research, with built in correlative studies based on less classical endpoints than those described above. Even so, *a priori* hypotheses (including sample size calculations) should be elaborated for these endpoints to guarantee the reliability of the analyses performed at the conclusion of the trial.

When are the endpoints compared?

In any trial of long duration, it is desirable to monitor the results in order to stop the trial early should any of the following circumstances occur:

A. an unexpected and serious toxicity has been observed
B. an unexpectedly large treatment difference has emerged, and continued accrual into the trial would be unethical
C. absolutely no therapeutic advantage has been observed, and continued accrual into the trial would be unnecessary.

These circumstances may be noticed at *interim analyses* of the trial data, which are often carried out once a year. The results of interim analyses must be interpreted with caution, even if they appear to be quite extreme, because the play of chance may cause a temporary difference in the results that will not hold in the longer run. Interim analyses can be planned in the protocol through use of sequential designs, or group sequential designs [18,19]. In essence, all these designs allow the trialists to analyse the data multiple times while controlling the probabilities of an erroneous conclusion.

In trials of substantial size, duration or importance, an Independent Data Monitoring Committee (IDMC)

Table 5
Main methods of analysis for phase III cancer clinical trials

Purpose of analysis	Nature of endpoint		
	Normal (e.g. white blood cell counts)	Binary (e.g. tumour response)	Time-dependent (e.g. survival)
Estimation	mean (and 95% c.i.)	proportion (and 95% C.I.)	median (and 95% C.I.) or Kaplan–Meier curves
Hypothesis test (unadjusted)	t-test or Wilcoxon non-parametric test	χ^2 test or Fisher exact test	logrank test
Hypothesis test (adjusted for covariates)	analysis of variance	Mantel–Haenszel χ^2 test	stratified logrank test
Regression analysis (with covariates)	linear regression model	logistic regression model	Cox regresssion model

C.I., confidence interval.

or Data and Safety Monitoring Board (DSMB) is often appointed to assess the interim results and to advise about the desirability to stop or to continue patient accrual [20]. The main advantage of an independent committee is to keep investigators blinded to the interim results of the trial, thereby avoiding any bias they could create, consciously or unconsciously, if they were privy to these results.

How are the endpoints compared?

The choice of an appropriate method of statistical analysis is obviously crucial. The number of methods that are commonly used in randomised clinical trials is, however, fairly limited. Table 5 shows the main methods for the analysis of normal, binary, or time-dependent endpoints. These methods are described in all standard statistical analysis packages.

How are treatment differences expressed?

It is essential, when reading a paper reporting the results of a phase III clinical trial, to identify the scale on which the treatment effect is expressed. Several scales are available, and each has pros and cons. Let us illustrate the choice of a scale on a made up example (Table 6). Suppose the outcome of interest in this trial is to reduce the incidence of an untoward event, from 50% in the control group to some lower percentage in the treated group.

Table 6
Hypothetical data showing the effect of treatment on an untoward event, and various measures of treatment effect

	Treated	Control
With event	45	50
Without event	55	50
Total	100	100
Risk of event	0.45	0.50

Absolute risk reduction = 5%; relative risk reduction = 10%; relative odds reduction = 18%; number needed to treat = 20.

The *absolute risk difference* is equal to the difference in the risks of the event in the two treatment groups: in the example of Table 6, $0.45 - 0.50 = -0.05$, i.e. an absolute risk reduction of 5%. The *relative risk* is equal to the ratio of the risks of the event in the two treatment groups: in the example of Table 6, $0.45/0.50 = 0.90$, i.e. a relative risk reduction of 10% ($= 1 - 0.90$). The *odds ratio* is equal to the ratio of the odds of the event in the two treatment groups: in the example of Table 6, $(0.45/0.55)/(0.50/0.50) = 0.82$, i.e. an odds reduction of 18% ($= 1 - 0.82$).

Note from Table 6 that the odds reduction is larger than the risk reduction, which in turn is larger than the absolute risk reduction. This is not a feature of the particular figures chosen in Table 6, it is a general feature that holds true for any treatment effect (other than zero). This fact should be kept in mind when reading a paper, and more importantly when comparing the results of different papers, since these may be expressed on different scales. It has been shown repeatedly that the same therapeutic benefit would lead to different prescription patterns depending on the scale used to express it, because any benefit seems more impressive when expressed in relative, rather than absolute, terms [21].

Some readers find it hard to have an intuitive grasp for risk reductions, whether absolute or relative. Treatment effects can also be expressed in terms of the *number needed to treat*, defined as the number of patients that must be treated for one untoward event to be avoided (on average). The number needed to treat is calculated simply as the inverse of the absolute risk reduction (without sign): in the example of Table 6, the number needed to treat is equal to $1/0.05 = 20$, i.e. on average 20 patients must be treated for one event to be avoided (which should not be taken as meaning that 19 patients do not benefit!)

Trial conduct

A poorly designed trial is likely to fail to answer the question it addresses. However, even a well-designed trial can fail if it is poorly conducted. It has therefore become standard practice to include, in the trial protocol, several sections describing the procedures used to conduct the trial. These procedures cover issues such as:

- the reporting of "Serious Adverse Events", and more generally of all adverse events observed over the course of the trial ("pharmacovigilance")
- the quality control system put in place to ensure the collection and processing of good quality data
- the adherence to the principles of Good Clinical Practice
- the quality assurance of the treatment procedures (doses, fractionation, schedules, compliance, etc.)
- the policies regarding publication and reporting of the trial results
- the legal sponsorship and the distribution of responsibilities between the various partners in the trial
- the informed consent process and the information for the patient, developed on the basis of national and local requirements.

References

1 ATAC Investigators. Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomized trial. Lancet 2002, 359: 2131–2139.

2 Buyse M. Centralized treatment allocation in comparative clinical trials. Applied Clin Trials 2000, 9: 32–37.

3 Hulstaert F, Van Belle S, Bleiberg H, et al. Optimal combination therapy with tropisetron in 445 patients with subtotal control of chemotherapy-induced nausea and emesis. J Clin Oncol 1994, 12: 2439–2446.

4 Citron ML, Berry DA, Cirrincione C, et al. Randomized trial of dose dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. J Clin Oncol 2003, 21: 1431–1439.

5 O'Connell MJ, Laurie JA, Kahn M, et al. Prospective randomized trial of postoperative adjuvant chemotherapy in patients with high-risk colon cancer. J Clin Oncol 1998, 16: 295–300.

6 O'Brien SG, Guilhot F, Larson RA, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. N Engl J Med 2003, 348: 994–1004.

7 Buyse M, Piedbois P. Should Dukes' B patients receive adjuvant chemotherapy? A statistical perspective. Seminars Oncol 2001, 28 (Suppl 1): 20–24.

8 Betensky RA, Louis DA, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomised clinical trials. J Clin Oncol 2002, 20: 2495–2499.

9 Machin D, Campbell MJ, Fayers PM, Pinol APY. Sample Size Tables for Clinical Studies. Blackwell Science Ltd, 1997.

10 Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. Lancet 1998, 352: 930–942.

11 Assmann AF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000, 355: 1064–1069.

12 Elsaleh H, Joseph D, Grieu F, Zeps N, Spry N, Iacopetta B. Association of tumour site and sex with survival benefit from adjuvant chemotherapy in colorectal cancer. Lancet 2000, 355:1745–1750.

13 Liver Infusion Meta-Analysis Group. Portal vein infusion of cytotoxic drugs after colorectal cancer surgery: a meta-analysis of 10 randomized studies involving 4000 patients. J Natl Cancer Inst 1997, 89: 497–505.

14 Simon R. Statistical tools for subset analysis in clinical trials. In: Scheurlen H, Kay R, Baum M, eds. Recent Results in Cancer Research. Heidelberg: Springer Verlag 1988; 111: 55–66.

15 DiLeo A, Bleiberg H, Buyse M. Overall survival is not a realistic endpoint for clinical trials in advanced solid tumors: A critical assessment based on recently reported phase III trials in colorectal and breast cancer. J Clin Oncol 2003, 21: 2045–2047.

16 Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P, for the Meta-Analysis Group In Cancer. Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Lancet 2000, 356: 373–378.

17 Bloom JC, Dean RA (Editors). Biomarkers in Clinical Drug Development. New York: Marcel Dekker, 2003.

18 Jennison C,Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC Press, 1999.

19 Whitehead J. The Design and Analysis of Sequential Clinical Trials (Revised second edition). Chichester: Wiley, 1997.

20 Ellenberg S, Fleming TR, DeMets D. Data Monitoring Committees in Clinical Trials: A Practical Perspective. New York: Wiley, 2002.

21 Tannock IF. From evidence-based medicine to clinical practice: not always straightforward. Eur J Cancer (this volume).