

Potential pitfalls in the design and reporting of clinical trials

Matthew R Sydes, Ruth E Langley

Lancet Oncol 2010; 11: 694–700

Published Online

May 26, 2010

DOI:10.1016/S1470-2045(10)70041-3

Cancer Group, Medical Research Council Clinical Trials Unit, London, UK (M R Sydes MSc, R E Langley MBBS); and Brighton and Sussex University Hospitals Trust, Brighton, UK (R E Langley)

Correspondence to:

Matthew R Sydes, Medical Research Council Clinical Trials Unit, 222 Euston Road, London NW1 2DA, UK
matthew.sydes@ctu.mrc.ac.uk

Randomised controlled trials are the gold standard method for developing evidence-based medicine. Good trial design and an awareness of some potential pitfalls are likely to maximise the chances of a successful trial with a conclusion that adds meaningfully to the evidence base. This paper is aimed at people early in their research careers and focuses on some common, usually avoidable, pitfalls in trial design. The areas covered include: assessing the scientific idea; trial design; size and duration of the trial; analysis; and reporting and presentation.

Introduction

Randomised controlled trials (RCT) are the gold standard method for developing evidence-based medicine. A good idea, strong belief in a therapy or clinical approach, and preliminary evidence from early trials are only the beginnings of a successful clinical trial. Although this paper, which is especially aimed at clinicians and research staff early in their research careers, focuses on phase 3 clinical trials in oncology, many of the issues raised will be relevant to other designs and disciplines.

Academic or investigator-initiated research is often supported by government or charity sources. Applications must be highly competitive to pass critical peer review and secure limited grant money. Regardless of whether the main result is positive or negative, recipients of grants are obliged to run high quality RCTs, the results of which are suitable for publication in a high-profile journal.

From the initial idea to the publication of the main findings of a trial, many changes can occur, including developments in clinical practice, new or competing products or approaches, and changes in the administrative and legal environment. Although some events are not predictable, good trial design and an awareness of some potential pitfalls are likely to keep the chances of success to a maximum. This paper focuses on five areas: assessing the scientific concept; trial design and development; size and duration of the trial; analysis; and reporting and presentation. A summary of the key issues is shown in the panel.

Assessing the scientific concept

Evaluating idea

Phase 3 trials compare new therapies or strategies with the current standard of care. To convince both clinicians and patients to participate, credible early evidence is required, sufficient to produce clinical uncertainty (equipoise). Evidence shows that phase 2 studies, which are often single-centre, non-randomised assessments with early outcome measures, such as response rate, commonly do not predict well later outcomes in the phase 3 setting.^{1,2}

Undertaking a formal systematic review before applying for grant funding is useful to help assess the idea, and can contribute to convincing funding bodies, peer reviewers, and potential investigators that the trial is worthy of support. This review should include relevant phase 2 and 3 trials and pertinent data—eg, toxic effects from observational studies. For example, when assessing a new idea to use transcutaneous oestrogen patches for prostate cancer, a formal systematic review was done.³ The review, which showed that intramuscular preparations of parenteral oestrogens had been used successfully in Scandinavia, significantly contributed to the scientific rationale for the PATCH trial (ISRCTN70406718). Funders might also consider trial proposals in the context of the National Institute of Health and Clinical Excellence (NICE) guidelines on current treatment and national research priorities.

Positioning the trial

An appreciation of ongoing national and international trials, including when they are likely to report and their possible effect on the current standard of care, is needed, especially if another trial might show the intended control arm to be inferior to new or other commonly used

Panel: Summary of issues

- Assessing the scientific concept: underestimating support or failing to appreciate ongoing work and its potential effect can lead either to a failure to secure funding or to poor recruitment, early closure, and inconclusive results. International collaboration might be required to address questions in a reasonable timeframe, but developing such collaborations can be complex and requires knowledge of the local regulatory and legal environment.
- Trial design and development: unambitious trial design can lead to results that add minimally to the evidence-base and do not best use the resources available. Although it is not possible to predict the entire course of a trial, anticipation of issues that might limit interpretation, such as narrow inclusion criteria, poorly defined outcome measures, and rigid stopping rules should be addressed in the design stage of the trial. Careful trial design is crucial to ensure that results will be pertinent when the trial reports, often many years after inception.
- Size and duration of trial: sample size calculation is based on several assumptions. Inaccurate or unrealistic estimates can result in failure to recruit or failure to analyse as planned. Overlong recruitment or follow-up leads to delays in publication of the findings, by which time clinical practice might have moved forward and the results will be considered obsolete.
- Analysis: analyses should be prespecified and documented. Failure to justify deviations from the planned analyses or a lack of clarity is likely to reduce the trial's credibility and can prohibit publication in an appropriate journal. A summary of the formal statistical analysis plan should be included in the trial protocol.
- Reporting and presentation: a key issue in presenting results is clarity so that readers can understand how the conclusions were reached. CONSORT diagrams and other tools can aid with this. A balanced, objective picture should be formed, which explains how the trial contributes to the current evidence-base.

treatments. A regimen known to be less efficacious than the potential control treatment should never be chosen as the control arm solely to show a larger effect size for the research arm. The increasing acceptance by the research community that ongoing clinical trials should be registered on publicly-accessible databases,⁴ such as ClinicalTrials.gov, aids with keeping track of ongoing trials and should avoid duplication and promote collaboration. If there are overlapping trials, investigators should be encouraged to plan pooled analyses of results in the future.

Trials are commonly developed by a "Trial Development Group". This group should have representation from the relevant clinical and statistical specialities, with members having a combination of time, energy, enthusiasm, and experience. Only concepts that are relevant and rigorously presented will be successful. Trial units have become fundamental to the conduct of successful trials, not least for their understanding of the regulatory environment. They also have expertise in trial design, statistics, and systematic reviews, established working relations with potential collaborators, and, often, experience with international collaboration. In the UK, there has been massive restructuring of clinical research, in cancer and other areas. Disease-specific clinical studies groups have the remit of identifying and prioritising important questions; therefore, engagement with these groups is essential from the outset. Trials adopted by these groups can access resources, such as research staff at sites, to aid with recruitment and data return.

Estimating support

A good estimate of the likely support for a trial is important. A formal survey can assess enthusiasm for a trial and provide feedback to refine the trial design. Estimates of predicted recruitment from sites should be treated with caution, because over-estimation is common; anecdotally, predicted recruitment is often divided by two to four. Estimates can be cross-checked against national registries to assess the number of patients potentially available.

Engaging representatives of patient groups at an early stage is recommended. They provide the potential participants' perspective and help ensure that a trial should be acceptable to patients. Some funding bodies have embraced this; for example, consumer involvement is strongly encouraged by Cancer Research UK. There are many disease-specific and generic patient groups, such as the James Lind Alliance and UK National Health Service (NHS) Involve.

International collaboration

International collaboration might be needed to accrue sufficient patients in a sensible timeframe, especially in less common tumour types or when postulations are relevant only to particular subgroups. International collaborations can be difficult to initiate. Involving potential international collaborators at an early stage in trial development allows broad contribution to trial

design and ensures compatibility with each country's practice and regulatory environment.

RADICALS, a trial assessing postoperative strategies for prostate cancer, has been launched by the Medical Research Council (MRC) and National Cancer Research Institute in the UK and the National Cancer Institute of Canada Clinical Trials Group.⁵ Both groups were closely involved in designing the trial, allowing it to open contemporaneously in each country. Neither country alone could realistically achieve the recruitment target of around 4000 patients over 5–6 years.

Clinical trials are required to have a sponsor that takes ultimate responsibility for trial management and conduct, including ensuring compliance with good clinical practice guidelines and pharmacovigilance. Differences in clinical trial regulations across national boundaries have made some sponsors unenthusiastic to sponsor outside of their own country.⁶ Negotiations with trial groups, rather than individual international sites, aids collaboration, because these groups will have knowledge of local legal and regulatory frameworks and might accept delegated responsibility for some of the sponsor's roles in their country. The collaborative Gynaecological Cancer Intergroup and European Study Group for Pancreatic Cancer are good examples of trial groups where trialists in many countries have successfully worked together on international trials.^{7–9}

Trial design and development

Recent advances in molecular and cellular biology, especially intracellular signalling pathways, growth factors and receptors, have led to a dramatic increase in the number of agents available that might improve disease outcomes. However, limited resources are available to undertake clinical trials, not only in terms of finance, but also in the number of patients willing to participate and in researchers' time.¹⁰ A realistic estimate of the time needed to initiate, undertake, and analyse a phase 3 trial in oncology is 5–10 years; thus, it is important to consider whether results will still be relevant after a decade and whether more than one scientific question can be addressed. Clinicians are often, understandably, cautious, about designing a trial that addresses more than one question, because of the increased complexity, especially in anticipation of explaining many treatment arms to potential patients. However, recent multi-arm trials, such as ICON5 (ISRCTN41636183),¹¹ FOCUS (ISRCTN79877428),¹² and STAMPEDE (ISRCTN78818544),¹³ have recruited excellently in ovarian, colorectal, and prostate cancer, respectively, suggesting that the approach is acceptable to patients and feasible in a busy clinical environment.

Design

Traditional phase 2 studies do not always predict phase 3 results.¹² Embedding a randomised phase 2 component into the initial stages of a phase 3 trial increases the amount of phase 2 data available to inform the design of

For more on the **James Lind Alliance** see <http://www.lindalliance.org>

For more on the **UK National Health Service (NHS) Involve patient group** see <http://www.invo.org.uk>

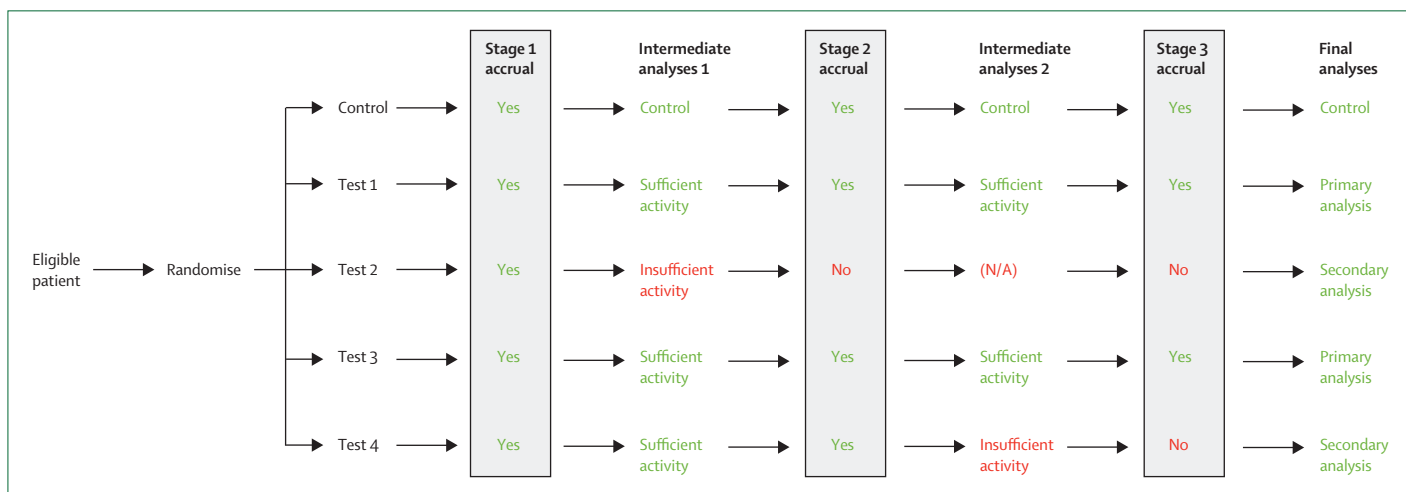


Figure 1: Example of a multi-group multi-stage trial

This hypothetical multi-arm multi-stage trial plans to assess five treatment groups (one control and four research) over three stages. All groups recruit at the start of the trial. At the intermediate and final analyses, each test group is compared in a pairwise fashion with the control group. After intermediate analyses, recruitment is stopped early to test groups, which are not showing sufficiently encouraging early evidence of activity; follow-up continues for patients already in those groups; all test groups are considered in the final analyses. In the example shown, test group 2 stops recruitment after the first intermediate analyses and test group 4 stops recruitment after the second intermediate analyses. In this way, resources are adaptively focused on the two test groups that look most promising (test group 1 and test group 3) and the control group.

the phase 3 component, offsetting some problems in estimation. Furthermore, patients who participate in the phase 2 component can potentially be included in the final phase 3 analysis. This results in fewer patients being required to address a clinical question and reduces the time needed, because there is no pause in recruitment between the phase 2 and 3 elements.

Statistical methods supporting trials that address more than one question concurrently include the factorial approach and the newer multi-arm, multi-stage (MAMS) design.¹⁴ Factorial designs assume that the effects of research arms are independent and the effects will be additive when combined. If this is not so and interactions do exist, the reported effect for each treatment can be overestimated or underestimated. Interactions manifest as synergy or antagonism between the treatments, where the effect of one treatment is larger or smaller in the context of the other treatment, or a ceiling effect where a second treatment cannot further improve outcomes.

MAMS designs separately compare a number of research arms against a common control arm in a seamless, phase 2 and 3 assessment, and adaptively focus on promising arms. The early stages focus on safety and early evidence of activity. The latter stages focus on efficacy. Recruitment is halted to arms not showing early evidence of sufficient benefit on an intermediate outcome measure—eg, response rate or progression-free survival—although follow-up continues for all patients and all research arms are considered in the final analyses.¹⁴⁻¹⁶ Recruitment can also be stopped early for safety in the usual way. Figure 1 shows a hypothetical MAMS trial assessing five arms (one control arm and four research arms) over three stages. At the intermediate and final analyses, each research arm is

compared pairwise to the control arm. In this way, the recruitment resources are adaptively focused on the most promising arms.

Population

RCTs aim to improve clinical outcomes in a given population. This population should be clearly defined in the trial's eligibility criteria. Inclusion or exclusion criteria that are unduly rigorous, especially if they exclude those with significant comorbidity or who are above a certain age, are likely to slow recruitment and limit the generalisability of the findings to the population of interest. Explanatory and pragmatic elements need to be balanced: explanatory designs assess treatments in more idealised settings, whereas pragmatic designs have greater applicability.¹⁷ An explanatory trial can be designed around a specific population to ensure relevance. For example, although frail, elderly patients with metastatic colorectal cancer are often treated with chemotherapy, but they are under-represented in most clinical trials.¹⁸ The FOCUS2 trial (NCT00070213) was specifically designed for this patient group, assessing chemotherapy regimens that start with lower doses and using outcomes specifically aimed at this group of patients.¹⁹

Credibility and consistency

Design issues that can affect the interpretation of findings should be considered prospectively. For phase 3 trials, treatment allocation must be assigned randomly by use of an acceptable method: there are very few reasons not to use central, computer-based allocation methods. In larger trials, randomisation alone should suffice, but stratification for known, important, prognostic factors will help control the balance of

treatment allocation, add credibility, and aid with subgroup analyses defined by these strata.

Primary and secondary outcome measures should be clearly defined, preferably by use of standardised, accepted definitions. In oncology, the Response Evaluation Criteria In Solid Tumours (RECIST), which define response and progression, are a good example.²⁰ However, other criteria are often ill-defined; for example, would a death caused by toxicity from chemotherapy count as a death-from-disease event in analyses of cause-specific survival? This issue is even more pertinent in subjective outcome measures, such as patient-reported outcome measures and quality-of-life scales. The primary outcome measure should always be clear; this should be identified and maintained throughout the trial. In uncommon instances, it is possible to revise the primary outcome measure, but this must be done with transparency and an awareness of how this will affect the trial's credibility and the perception of these findings.

Formal interim stopping rules should be defined carefully if they are used. There are many examples of trials stopping early after a perceived advantage in an intermediate outcome measure with later analyses showing no evidence of a benefit in the definitive measure or the estimates of intermediate measures having regressed towards the mean from a random high.²¹ A more flexible approach is the lack-of-benefit approach, which ensures a trial will stop either if there is clear evidence of harm—eg, unacceptable toxic effects—or an absence of sufficient benefit, based on predefined guidelines.²² This approach, used in the MAMS and other designs, considers only data that have been collected so far and differs from futility monitoring, which must make assumptions about the probable nature of future data.²³ Trials should not be stopped early unless there is sufficient evidence to convince most clinicians that the trial's question has been answered. In view of the fact that a large proportion of trials do not produce positive results, wider use of lack-of-benefit analyses would allow earlier investment into other potentially beneficial approaches.

Size and duration of the trial

Determining sample size is a key design issue. Two factors especially need clinical input: estimating the control arm event rate and determining a clinically meaningful difference to target between the control and investigational therapy. The control arm event rate is the point from which the difference between arms is calculated and will strongly affect the number of patients needed. Accrual rates and follow-up duration are other important parameters.²⁴ It is important to remember that it is the number of events (endpoints) that determines power, not the number of patients; the number of patients needed is back-calculated from the events required. The effect on sample-size calculations of misestimating the control arm event rate is shown in table 1. If the event rate is lower than predicted, there will be fewer events (thus less power) at the planned

	4-year survival		Accrual over 3 years (n)	Events after 5 years (n)	Power after 5 years (%)
	Control arm (%)	Research arm (%)			
Reference	60%	68%	1196	383	80%
Higher survival—same accrual	70%	77%	1196	284	67%
Higher survival—same power	70%	77%	1618	383	80%
Higher survival—same accrual	80%	85%	1196	187	50%
Higher survival—same power	80%	85%	2457	383	80%
Lower survival—same accrual	50%	60%	1196	485	88%
Lower survival—same power	50%	60%	942	383	80%

Accrual rate is presumed to be constant and the target hazard ratio is maintained at 0.75. The reference scenario sets the trial to recruitment for 3 years in order to have 80% power to detect a 25% reduction in the risk of death from 60% to 68% at 4 years. Table 1 shows what the observed power would be at 5 years if the event rate differed from the reference scenario or what the necessary accrual rate over 3 years would be to attain 80% power at 5 years. For example, to see a 25% reduction in risk (hazard ratio 0.75) in 4-year survival from 60% to 68% with 3-year accrual and analyses planned for 2 years later would require that 1196 patients be randomised over those three years. However, if the event is lower such that 4-year survival in the control arm is 70%, the trial would have only 67% power at 5 years, instead of 80% power. To have 80% power at 5 years, the trial would have had to recruit 1618 patients over those 3 years.

Table 1: Effect of misestimating event rates on accrual and power

	4-year survival		Hazard ratio	Accrual over 3 years (n)	Events after 5 years (n)
	Control arm (%)	Research arm (%)			
Superiority trial (80% power with a two-sided 5% significance level)					
Scenario 1	60	65	0.84	3247	1084
Scenario 2	60	70	0.70	791	247
Scenario 3	60	75	0.56	340	99
Scenario 4	60	80	0.44	184	49
Non-inferiority trial (95% power with a one-sided 5% significance level)					
Scenario 1	65	60	1/0.84	4346	1451
Scenario 2	70	60	1/0.70	1059	330
Scenario 3	75	60	1/0.56	456	132
Scenario 4	80	60	1/0.44	246	66

The example fixes accrual for 3 years, assuming a constant accrual rate, and trial duration to be 5 years. Superiority trials aim to show a difference at least as large as the hazard ratio shown; the non-inferiority trials aim to exclude a difference as large as the hazard ratio shown. For example, to demonstrate superiority from 60% to 70% in 4-year survival (hazard ratio 0.70) in a 5-year trial with 3-years recruitment would require 791 patients be randomised. If control arm 4-year survival is 70% then to rule out a hazard ratio of 1.43 (1/0.70) and research arm 4-year survival of 60% in a 5-year trial with 3-years recruitment would require 1059 patients to be randomised.

Table 2: Effect of changing the target difference on accrual

analysis time. Extending recruitment or delaying analyses would increase the number of events (and power), but this might extend the trial beyond the funding grant period or the results might be less relevant when published.

A larger target difference requires fewer events, so there is commonly a temptation to be optimistic, about the potential success of a research therapy. Table 2 shows how the number of patients required decreases when the target difference increases. A smaller difference between treatments requires more events to determine the true effect size. Recent improvements in clinical outcomes in oncology reflect a series of modest improvements rather than new, so-called wonder therapies, with few notable exceptions.^{25,26} Small differences might not be seen as

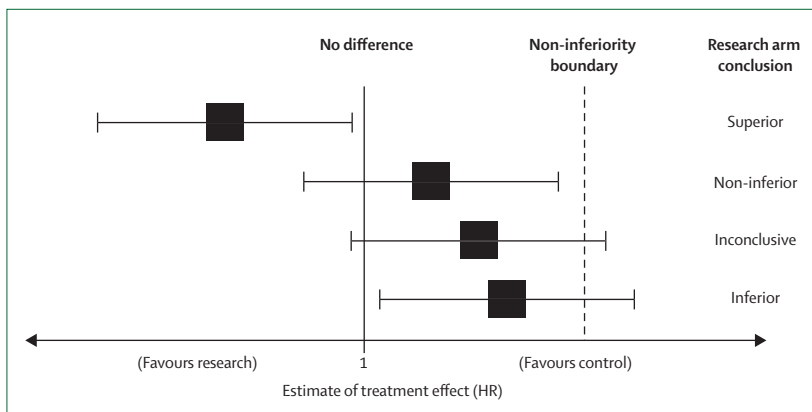


Figure 2: Conclusions from estimated treatment effect (hazard ratio) in non-inferiority trial
 Lines show estimate of treatment effect and confidence interval. HR<1 favours research group; HR>1 favours control group. HR=hazard ratio.

sufficiently worthwhile to change clinical practice, especially if the new therapy is costly, complicated, or toxic. Planning for modest improvements is probably more realistic—eg, a hazard ratio, in time-to-event analyses, in the range 0.75–0.80, translating to 20–25% relative reduction in risk and an absolute improvement of 4–10%, depending on the control-arm event rate.

The choice of a superiority or non-inferiority design is an important decision affecting trial size. A superiority trial is designed to determine whether the research treatment is better than the control treatment in terms of efficacy. In a non-inferiority trial, the aim is to show that the research treatment is not worse than the standard approach in some clinically important regard;²⁷ this is done by defining a “lower” border to the relative efficacy, which would be considered an acceptable detriment. (An equivalence trial defines both a lower and an upper border of relative efficacy, within which the treatments would be considered the same.) Non-inferiority recognises that some detriments would be balanced by other expected advantages. Figure 2 shows some potential conclusions from a non-inferiority trial. A common belief is that non-inferiority trials will be simpler to do; the converse is often true, because the acceptable detriment (range of non-inferiority) tends to be small, usually smaller than the minimum benefit needed for a superiority trial, and so needs a larger number of events and, thus, more patients (table 1). If non-inferiority is the aim, it is important to document how the non-inferiority boundary was defined and who considered this acceptable,²⁸ especially noting any consumer involvement.

Analysis

Analyses should be prespecified, expressed in a formal statistical analysis plan (ICH E9), written by the statistician with clinical input, and have a summary or first formal version included in the protocol. This enhances the trial’s credibility and allows researchers to be guided by their previous plans rather than driven by interesting, but potentially spurious, patterns in the data. Other analyses should be considered exploratory and labelled as such.

Particular care should be taken with subgroup analyses. Results of subgroups are sometimes analysed in subgroups defined after reference to the data, referring only the subgroup in which an effect was seen and without the use of interaction tests. Therefore, it is possible to wrongly identify subgroups. If unplanned subgroup analyses are included, they should be justified and, usually, treated as hypothesis-generating analyses. Good guidance on this matter is available.^{29–31}

Both the planned and the observed number of events for the primary outcome measure should be reported to express the trial’s maturity. If there are fewer events than planned, one should consider how much follow-up information is available, whether the results are being reported prematurely, and whether increased follow-up would be likely to change the findings? It should be clear who has seen interim outcome data and how often the data have been reviewed.³² Repeating analyses of the same outcome increases the probability of obtaining, by chance alone, results that are considered “statistically significant”—eg, false-positive results—or “random highs” that are not real. Therefore, interim data should not be analysed too frequently and a meaningful number of events should accrue between analyses. Data should always be as current as possible for analyses; thus, the date the dataset was frozen should be reported. Follow-up assessments and rates of missing data should be similar on each arm. Sensitivity analyses exploring the effect of missing data on the findings and an investigation into whether the amount of missing data is affected by allocated treatment can be done.³³

In superiority trials, intention-to-treat (ITT) analyses are standard—ie, including all patients in the analyses according to the allocated treatment. This leads to a conservative estimate of effect size; however, by excluding patients who did not start treatment or who stopped during treatment leads to biased estimates. Exceptions to ITT analyses are sometimes permissible, but any modifications should be documented. For example, in the UKLG LY09 Hodgkin’s lymphoma trial,³⁴ patients were excluded from ITT analyses if postrandomisation central pathology review showed they did not have Hodgkin’s lymphoma. Central pathology review, independent of allocated treatment, was planned for all patients, but was not practicable before randomisation. Only 19 of 807 patients who proved to be ineligible on review were excluded from the main analyses; this modification of ITT was accepted by peer reviewers. If many patients are ineligible, other appropriate steps would be required, such as increasing the accrual target.

In per-protocol analyses, patients are analysed only if they have complied with the protocol in a prespecified fashion—eg, completed all treatment cycles. These analyses represent an idealised interpretation of the results if treatment can be given. Because it is not known in advance which patients will be able to tolerate treatment, care should be taken in extrapolating estimates of efficacy from per-protocol analyses to the every day clinical setting.

For more on ICH E9 see <http://www.ich.org/LOB/media/MEDIA485.pdf>

Conversely, per-protocol (“on-treatment”) analyses are usually recommended in non-inferiority trials, because they provide more conservative estimates, although are possibly still biased. “As-treated” analyses, in which patients are included according to the treatment they received, are generally more appropriate for safety and toxicity analyses, because the proportion of patients with side-effects can be underestimated if patients who do not start treatment are included or if patients who cross-over to another treatment are excluded.²⁴ The criteria for inclusion in the per-protocol analyses should be specified before the data are assessed and reiterated with each presentation.

Reporting and presentation

This section concentrates on the reporting of statistical analyses and the interpretation of these findings. Useful guidance for reporting of clinical trials is given in the Consolidated Standards of Reporting Trials (CONSORT) guidelines and its extensions.^{35,36}

Baseline characteristics of patients entered into the trial should be presented, but formal comparisons of baseline characteristics are not recommended: with successful randomisation, any imbalances in baseline characteristics can only have arisen by chance and can be accounted for in adjusted analyses.^{35–37} Any departure from the expected population should be noted—eg, under-representation of elderly patients. Caution should be taken in extrapolating findings beyond the population actually recruited.

For time-to-event outcome measures—eg, survival—it is better to use the hazard ratio, which compares the whole time-to-event experience, rather than assessing the difference between treatment arms only at a given timepoint. The degree of uncertainty surrounding, or reliability of, an estimate such as a hazard ratio should be expressed with a confidence interval (interval estimate) to aid interpretation, usually at the 95% level.

If the difference between the treatments at a given timepoint is required, using values from Kaplan-Meier plots can be unreliable, because, by chance, the curves can be closer or further apart in absolute terms than reflects the whole trial experience. It is better to determine the event-free proportion in the control arm and apply the overall HR to estimate the event-free proportion in the research arm. This estimate of the difference has only the unreliability of one Kaplan-Meier curve rather than two. This applies equally to estimates of median survival. Better yet, it is possible to use flexible parametric models to smooth the Kaplan-Meier lines to better reflect the patient experiences.³⁸ More reliable values can be read from these graphs with better estimates of the differences between treatments over time; a graphical example of this approach is provided in the paper by Dearnaley and colleagues.³⁹

As time from randomisation increases, Kaplan-Meier curves are typically based on fewer data and become less reliable; thus, the numbers of patients at risk should be presented at salient time points. The extreme right-hand end of Kaplan-Meier graphs should not be presented,

because the estimates are unreliable when there are few patients contributing. The number of events and confidence interval should be clear; for example, the first published report of the MRC RT01 trial³⁹ focused on the primary outcome measure, biochemical progression-free survival. The trial’s secondary, later outcome measures were also presented, including local clinical disease progression, where marked improvement was noted with a hazard ratio of 0·65; but only 32 local disease progression events had been reported from about 800 patients, so the confidence intervals were wide (0·36–1·18). Although the point estimate suggests a clinically important advantage to the research arm, the confidence interval doesn’t rule out an 18% relative detriment. Immature results such as this should not be overinterpreted; trial follow-up continues.

Sensationalist terms should be avoided and it should be clear whether differences are expressed in relative or absolute terms: a doubling of risk sounds large, but might be small in absolute terms—eg, an increase from 2% to 4%; although if a drug is used widely enough even a small absolute difference might have a marked effect on the population.⁴⁰ Percentages are always best presented together with the numbers from which they were calculated. A surprising number of graphs have problems involving the selection of tick marks on the axes. Graphs depicting the proportion of people who are event-free should have y-axes running between 0 (all patients with events) and 1 (no patients with events, yet). Truncating the axes to between 0·8 and 1·0, for example, can inappropriately exaggerate the apparent difference between the treatment arms. Presenting cumulative events starting at “none” obviates this problem.⁴¹

An appropriate balance should be given to both efficacy and safety findings and the results should be placed in the context of other known data. Where possible, a systematic review should be done, preferably updating a review undertaken at the design stage; if possible, these data should be combined in a meta-analysis to put the new results in the context of the accumulated evidence. The limitations of the trial should be acknowledged and what steps have been taken to minimise their effect.

Conclusion

Clinical trials are the backbone of evidence-based medicine and are likely to remain that way for the foreseeable future. Improved public engagement and public acceptance of clinical trials, through the efforts and openness of all researchers and increased lay understanding, is likely to lead to greater participation in clinical research and better use of resources. It is important that trials are designed

Search strategy and selection criteria

The examples used are taken mostly from our own experience and practice, but also from purposive search of the published work to highlight these issues and possible solutions.

efficiently, done well, and complement clinical practice. Trials must be reported clearly so that accumulated information can be learned and shared. The best way to avoid potential pitfalls is to acknowledge that they exist, be clear where they might be found, and to work to avoid them.

Contributors

MRS led the writing of the manuscript, which REL co-wrote, and both authors read and approved the final version.

Conflicts of interest

The authors declared no conflicts of interest.

Acknowledgments

The authors thank Neal Navani, Max Parmar, and Martha Perisoglou for their comments on drafts of the manuscript.

References

- Chen TT, Chute JP, Feigal E, Johnson BE, Simon R. A model to select chemotherapy regimens for phase III trials for extensive-stage small-cell lung cancer. *J Natl Cancer Inst* 2000; **92**: 1601–07.
- Freidlin B, Breathnach OS, Johnson BE. A model to select regimens for phase III trials for patients with advanced-stage non-small cell lung cancer. *Clin Cancer Res* 2003; **9**: 917–22.
- Norman G, Dean ME, Langley RE, et al. Parenteral oestrogen in the treatment of prostate cancer: a systematic review. *Br J Cancer* 2008; **98**: 342–51.
- De Angelis C, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Lancet* 2004; **364**: 911–12.
- Parker C, Clarke N, Logue J, et al. RADICALS (Radiotherapy and Androgen Deprivation in Combination after Local Surgery). *Clin Oncol* 2007; **19**: 167–71.
- Hearn J, Sullivan R. The impact of the 'Clinical Trials' directive on the cost and conduct of non-commercial cancer trials in the UK. *Eur J Cancer* 2007; **43**: 8–13.
- Poveda A. Ten years of "Optimal Therapy in Advanced Ovarian Cancer. Update" meeting. *Int J Gynecol Cancer* 2008; **18**: 67–70.
- European Study Group for Pancreatic Cancer. http://www.pancsoc.org.uk/new/index_en.php?pageid=8&subpageid=0 (accessed March 17, 2010).
- Gynecologic Cancer Intergroup. GCIG: Gynecologic Cancer Intergroup. Enhancing the global impact of clinical trials in gynaecologic cancer. <http://www.gcig.igcs.org/> (accessed March 17, 2010).
- Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009; **374**: 86–89.
- Copeland LJ, Bookman M, Trimble E. Clinical trials of newer regimens for treating ovarian cancer: the rationale for Gynecologic Oncology Group Protocol GOG 182-ICON5. *Gynecol Oncol* 2003; **90**: S1–S7.
- Seymour MT, Maughan TS, Ledermann JA, et al. Different strategies of sequential and combination chemotherapy for patients with poor prognosis advanced colorectal cancer (MRC FOCUS): a randomised controlled trial. *Lancet* 2007; **370**: 143–52.
- James ND, Sydes MR, Clarke NW, et al. STAMPEDE: Systemic Therapy for Advancing or Metastatic Prostate Cancer—a multi-arm multi-stage randomised controlled trial. *Clin Oncol (R Coll Radiol)* 2008; **20**: 577–81.
- Parmar MKB, Barthel F, Sydes M, et al. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst* 2008; **100**: 1204–14.
- Royston P, Parmar MK, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med* 2003; **22**: 2239–56.
- Sydes MR, Parmar MK, James ND, et al. Issues in applying multi-arm multi-stage (MAMS) methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* 2009; **10**: 39.
- Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials* 2009; **10**: 37.
- Aapro MS, Kohne CH, Cohen HJ, Extermann M. Never too old? Age should not be a barrier to enrollment in cancer clinical trials. *Oncologist* 2005; **10**: 198–204.
- Seymour MT, Maughan TS, Wasan HS, et al. Capecitabine (Cap) and oxaliplatin (Ox) in elderly and/or frail patients with metastatic colorectal cancer: the FOCUS2 trial. *Proc Am Soc Clin Oncol* 2007; **25**: 9030.
- Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; **45**: 228–47.
- Cannistra SA. The ethics of early stopping rules: who is protecting whom? *J Clin Oncol* 2004; **22**: 1542–45.
- Freidlin B, Korn EL. Monitoring for lack of benefit: a critical component of a randomized clinical trial. *J Clin Oncol* 2009; **27**: 629–33.
- Freidlin B, Korn EL. A comment on futility monitoring. *Control Clin Trials* 2002; **23**: 355–66.
- Girling D, Parmar M, Stenning S, Stephens R, Stewart L. Chapter 5: Trial size. In: *Clinical trials in cancer: principles and practice*, 1st edn. Oxford: Oxford University Press, 2003: 83–114.
- Roberts TG Jr, Lynch TJ Jr, Chabner BA. The phase III trial in the era of targeted therapy: unraveling the "go or no go" decision. *J Clin Oncol* 2003; **21**: 3683–95.
- Bailer JC III, Gornik HL. Cancer undefeated. *N Engl J Med* 1997; **336**: 1569–74.
- Fleming TR. Current issues in non-inferiority trials. *Stat Med* 2008; **27**: 317–32.
- Le Henanff A, Giraudeau B, Baron G, Ravaut P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA* 2006; **295**: 1147–51.
- Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007; **357**: 2189–94.
- Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**: 176–86.
- Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21**: 2917–30.
- DAMOCLES study group, NHS Health Technology Assessment Programme. A proposed charter for clinical trial data monitoring committees: helping them to do their job well. *Lancet* 2005; **365**: 711–22.
- Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.
- Johnson PW, Radford JA, Cullen MH, et al. Comparison of ABVD and alternating or hybrid multi-drug regimens for the treatment of advanced Hodgkin lymphoma: results of the UK Lymphoma Group LY09 trial (ISRCTN97144519). *J Clin Oncol* 2005; **23**: 9208–18.
- Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**: 1191–94.
- CONSORT group. CONSORT statement. <http://www.consort-statement.org/> (accessed March 17, 2010).
- Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990; **335**: 149–53.
- Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med* 2002; **21**: 2175–97.
- Dearnaley DP, Sydes MR, Langley RE, et al. The early toxicity of escalated versus standard dose conformal radiotherapy with neoadjuvant androgen suppression for patients with localised prostate cancer: Results from the MRC RT01 trial (ISRCTN4772397). *Radiother Oncol* 2007; **83**: 31–41.
- Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med* 2005; **352**: 1092–102.
- Pocock SJ, Trivison TG, Wruck LM. How to interpret figures in reports of clinical trials. *BMJ* 2008; **336**: 1166–69.