Strategy for randomised clinical trials in rare cancers

Say-Beng Tan, Keith B G Dear, Paolo Bruzzi, David Machin

Proving that a new treatment is more effective than current treatment can be difficult for rare conditions. Data from small randomised trials could, however, be made more robust by taking other related research into account

The need for randomised trials to establish that treatments are effective is well established. However, because the effects of new treatments are usually modest compared with standard treatment, large numbers of patients are needed to detect any genuine benefits. This means that, even for common cancers, studies often have to be multicentred to ensure enough patients are recruited in a reasonable time. The strategy for testing new treatments in rare cancers, where it is impossible to accrue large number of patients, is unclear. We extend Lilford and others' proposal that a bayesian statistical approach, using related information from earlier studies, would be useful in designing and subsequently summarising small randomised controlled trials.[1] We suggest a scoring system for pooling this evidence and detail how this may be combined with hypothetical scenarios to assist in the design of, and justification for, a small randomised controlled trial.

Summary points

- Treatments for rare cancers are difficult to evaluate in randomised trials as there are too few patients to detect any genuine treatment differences
- Combination of previous data with trial data by bayesian techniques could help overcome this problem
- Data from related studies are scored and weighted according to pertinence, validity, and precision
- The method of combining data increases the robustness of information from small trials and can be used to help design and provide justification for such trials

Problems of small trials

Randomised controlled trials are regarded as the standard when comparing a new treatment with the standard treatment for a particular cancer. However, to be considered clinically worth while in clinical trials, these (essentially) very toxic regimens typically need to show relative reductions in the risk of death of 20-30%. For studies to have sufficient statistical power ($\geq 80\%$) to detect treatment effects of this magnitude, several hundreds of deaths (typically 200 to 500) need to be observed. This implies trial sizes that are unrealistically large for rare cancers. Furthermore, even if a much larger treatment effect could be expected, estimates derived from the resulting (small) randomised controlled trial would lack the precision needed for clinical decisions.

Thus, investigators who wish to test new treatments for rare cancers tend to conduct either single arm studies of tumour response rates or comparative studies using historical controls. Alternatively, investigators may attempt to conduct small, underpowered, randomised controlled trials. These give rise to estimates of outcome that have unacceptably large confidence intervals and thus fail to provide clear answers. On these grounds, protocol review boards may regard such trials as unethical.[2] However, some have argued that in situations such as rare diseases, small randomised trials are the "only way that any unbiased measurements of effectiveness can be made."[3]

One suggested solution to the problem is to use bayesian statistical approaches.[4] These involve quantifying the information available about the outcome of interest in the form of a prior probability distribution at the design stage and combining this with the trial data to give a posterior distribution. Conclusions are then drawn from the posterior distribution. The key step in a bayesian approach is summarising the information available before the trial. This will often be from single arm studies or studies of response rate rather than survival.

Designing a new trial

Suppose we wish to design a randomised controlled trial to compare a new treatment with the standard treatment for a rare cancer, with the primary endpoint being overall survival. In such a case we would typically estimate the corresponding survival curves by the Kaplan-Meier technique and use the hazard ratio to estimate the magnitude of the treatment difference.[5] It is customary when designing any trial to summarise the available information related to the question under

Division of Clinical Trials and Epidemiological Sciences, National Cancer Centre, 11 Hospital Drive, Singapore 169610
Say-Beng Tan *senior biostatistician*

National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia
Keith B G Dear *senior fellow*

Unit of Clinical Epidemiology and Trials, National Cancer Research Institute, Genoa, Italy
Paolo Bruzzi *head*

United Kingdom Children's Cancer Study Group, University of Leicester, Leicester
David Machin *professor*

Correspondence to:
S-B Tan
ctetsb@nccs.com.sg

BMJ 2003;327:47–9

P+

A paper giving a worked example of the method is available on bmj.com

Table 1 Proposed scales and scores for assessing the three components of pertinence of study relevant to small randomised controlled trial under design

Cancer	Treatment	Endpoint	Component score
Same disease and stage	Same as proposed standard and experimental treatments	Overall survival	1
Same disease, different stage or type of patient	Same standard treatment, similar experimental treatment (eg different dose)	PFS, DFS, or EFS; adjustment factor available	0.9
Different site, same biology/histology	Similar standard and experimental treatments	PFS, DFS, or EFS; adjustment factor unavailable	0.8
Same site, different biology/histology		Response rate validated as a surrogate endpoint	0.5
Different site, some similarity	Some similarity in standard or experimental treatment, or both	Response rate not validated as surrogate end point	0.3
Different disease	Unrelated treatments	Unrelated end points	0

DFS=disease-free survival, PFS=progression-free survival, EFS=event-free survival.

consideration and to specify what would be regarded as a clinically important difference between the treatments being compared. In addition, there is a need to specify the type of patients eligible for the trial and quantify the number of patients that are likely to be recruited in a reasonable time frame.

Summarising and weighting information

For our model we assume that the evidence available at the planning stage of the proposed trial is from several studies, each providing relevant information according to three main criteria: pertinence, validity, and precision.

Pertinence

Pertinence summarises how close the information is to that which we wish to obtain during the proposed trial. The component parts to a full assessment of pertinence are the precise cancer investigated, the treatment(s) evaluated, and the endpoint measure. Table 1 lists six pertinence levels for these components with a score from 0 (no pertinence) to 1 (fully pertinent) for each. The minimum of the component scores provides an overall pertinence score (PS) for each study. By using the minimum score, we fully acknowledge whatever is the most serious defect of each study. An alternative would be to use the product of the scores, but this gives pertinence greater influence than validity (see below), which has only one component.

The adjustment factor (table 1) enables hazard ratios calculated from, for example, event-free survival to be converted to hazard ratios for overall survival when these are not reported. The adjustment factor is calculated as the ratio of two hazard ratios obtained from other studies that report ratios for both event-free survival and overall survival. If other end points are quoted, their relevance will depend on whether they have been validated as a surrogate for overall survival.[6 7]

Validity

Validity measures the quality of the available studies and depends on their design. It is maximal for properly designed and conducted randomised controlled trials and minimal for case reports. Table 2 gives a proposed classification and suggested validity scores.

Precision

Precision indicates how reliably the hazard ratio is determined. It depends on the number of events reported in each group. The more events there are, the more precise the ratio. Note that the study specific hazard ratio should be obtained, even from single arm studies without a control group. If necessary, this can be obtained by comparing the results with those from historical control group(s) mentioned in the study report or on some other clearly explained basis.

Correction factor

Once each study has been scored, the precision and validity scores are used as a correction factor to down weight the information. This is done by multiplying the number of events by the two scores in turn. This adjusted number of events reflects the added uncertainty associated with the methodological limitations of the study or with its limited pertinence to the question of interest. Other correction factors for the estimated hazard ratio may also be introduced at this stage—for example, to take into account the overestimate of treatment effects that is typically observed in uncontrolled studies.

Prior and posterior distribution

The adjusted numbers of events from each study are used to calculate the weighted mean prior log hazard ratio (LHR_{Prior}). The prior distribution is then constructed as a normal distribution with mean $\mu_{Prior}=LHR_{Prior}$ and standard deviation $\sigma_{Prior}=\sqrt{(4/m_{Prior})}$, where m_{Prior} is the adjusted number of events (deaths) from all the studies. If no prior information exists, a subjective prior distribution can be elicited from the investigators or other experts.[8–10]

If μ_{Data} is the log hazard ratio based on actual deaths (m_{Data}), then (following Parmar et al[11]) the posterior distribution has a normal distribution with mean $\mu_{Posterior}=(m_{Prior}\mu_{Prior+}m_{Data}\mu_{Data})/(m_{Prior+}m_{Data})$ and standard deviation $\sigma_{Posterior}=\sqrt{[4/(m_{Prior+}m_{Data})]}$. At the planning

Table 2 Proposed scores for assessing the validity of study relevant to small randomised controlled trial under design

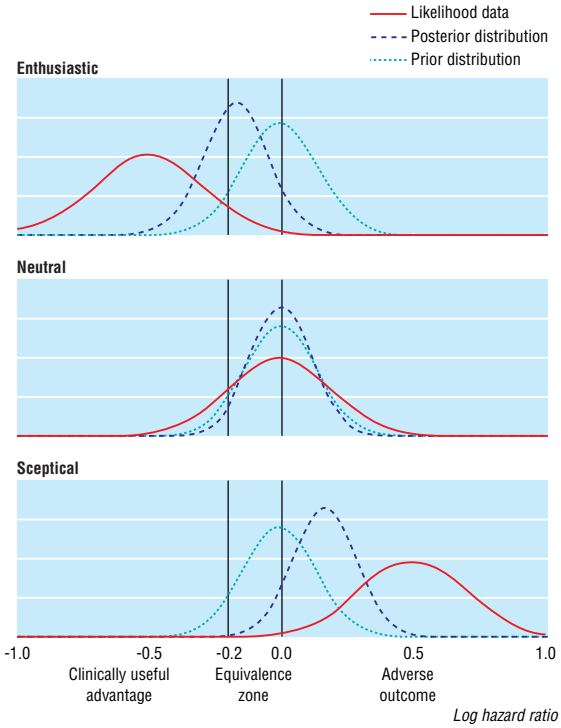
Design	Validity score
Randomised controlled trial:	
No major flaws	1
Questionable quality	0.8
Major flaws	0.6
Non-randomised trial:	
Prospective controlled	0.4
Single arm study:	
With prespecified historical controls	0.3
No historical controls	0.2
Case study:	
Series	0.1
Single report	0.05

stage, μ_{Data} and m_{Data} are obtained from hypothetical scenarios, but once the trial is completed they are obtained from the actual data.

Using scenarios

Once we have constructed the prior distribution and determined the number of patients likely to be recruited, the next step is to consider (at least) three hypothetical scenarios for the outcome of the trial: enthusiastic (experimental treatment is clearly better), neutral (treatment is the same), and sceptical (treatment worse than the control). These correspond to datasets with log hazard ratios that are negative, zero, or positive (figure). The prior distribution is combined with the hypothetical datasets to give a posterior distribution. For example, when the data from the enthusiastic scenario are combined with the prior distribution, which assumes no difference between treatments in this example, the final results suggest an almost 50% probability of a clinically useful advantage to the experimental treatment (see the area under the posterior distribution curve of the figure).

These scenarios can be presented to the protocol review board to help show that useful conclusions can be drawn from the proposed "small" trial. If the trial is given the go ahead, the posterior distribution is obtained from the real data combined with the prior distribution derived in the planning stage[12] or one modified by new information that becomes available during the trial.[13]



Prior and posterior distributions and likelihood "data" for enthusiastic, neutral, and sceptical scenarios. The peak of each distribution corresponds to the most likely value of the true log hazard ratio, as estimated from the prior studies or the "data" or both (see text). The prior distribution in this example assumes that the two treatments are identical (log hazard ratio=0)

Discussion

Our proposed approach offers one way to overcome the problem of evaluating new treatments for rare conditions in randomised trials. Inferences based on the posterior distributions obtained in the approach provide a fuller description of the improved state of knowledge than is available from merely quoting P values or confidence intervals. An example of how the model could be applied in practice for supratentorial primitive neuroectodermal tumour, a rare childhood cancer, is available in our companion paper (see bmj.com).

The pertinence and validity scores that we have suggested to grade the usefulness of prior evidence should be modified as necessary to suit each project, although the investigators should clearly state the scoring systems that they use. Other scoring systems have been proposed,[14 15] but a key feature of ours is that it reflects the relevance to a specific research question and is not simply a measure of overall quality. When all prior information derives from randomised trials, our approach reduces to a standard meta-analysis. When no suitable prior information is available, a distribution can be constructed by eliciting the opinions of experts.

We are not proposing an alternative to conducting trials of adequate size when these are feasible nor providing a justification for single centre studies where multicentre studies are clearly the best option. Rather, we hope that our proposals will contribute to the establishment of a clear strategy for randomised clinical trials in rare cancers and other conditions.

Contributors: KBGD, PB, and DM first considered and discussed the small trials problem. SBT and DM proposed the use of a bayesian approach and then outlined the initial approach to be taken. All authors contributed to the subsequent development of the methodology and the writing of the paper.

Funding: SBT is funded by the National Medical Research Council of Singapore

Competing interests: None declared.

1 Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of the conundrum. *BMJ* 1995;311:1621-5.
2 Edwards SJL, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. *Lancet* 1997;350:804-7.
3 Lilford R, Stevens AJ. Underpowered studies. *Br J Surg* 2002;89:129-31.
4 Spiegelhalter DJ, Myles P, Jones DR, Abrams KR. An introduction to bayesian methods in health technology assessment. *BMJ* 1999;319: 508-12.
5 Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. II: analysis and examples. *Br J Cancer* 1977;35:1-39.
6 Begg CB, Leung DHY. On the use of surrogate end points in randomized trials. *J R Stat Soc A* 2000;163:15-24.
7 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605-13.
8 Parmar MKB, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with bayesian methods. *Lancet* 2001;358:375-81.
9 Tan SB, Machin D, Cheung YB, Chung YFA, Tai BC. Following a trial that stopped early: what next for adjuvant hepatic intra-arterial iodine-131-lipiodol in resectable hepatocellular carcinoma? [letter]. *J Clin Oncol* 2002;20:1709.
10 Tan SB, Chung YFA, Tai BC, Cheung YB, Machin D. Elicitation of prior distributions for a phase III randomized controlled trial of adjuvant therapy with surgery for hepatocellular carcinoma. *Control Clin Trials* 2003;24:110-21.
11 Parmar MKB, Spiegelhalter DJ, Freedman LS. The CHART trials: bayesian design and monitoring in practice.. *Stat Med* 1994;13:1297-312.
12 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J R Stat Soc A* 1994;157:357-87.
13 Machin D. Discussion of "The what, why and how of bayesian clinical trials monitoring." *Stat Med* 1994;13:1385-9.
14 Smith MB, Feldman W. Over-the-counter cold medications. A critical review of clinical trials between 1950 and 1991. *JAMA* 1993;269:2258-63.
15 Jadad AR, Moore RA, Carrol D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports on randomized clinical trials: is blinding necessary? *Controlled Clin Trials* 1996;17:1-12.

(Accepted 30 May 2003)