BMJ

RESEARCH

Reporting of sample size calculation in randomised controlled trials: review

Pierre Charles, research fellow in epidemiology, specialist registrar in internal medicine,[1,2,3] Bruno Giraudeau, assistant professor of statistics,[1,4,5,6] Agnes Dechartres, research fellow in epidemiology,[1,2,3] Gabriel Baron, statistician,[1,2,3] Philippe Ravaud, professor of epidemiology[1,2,3]

[1]INSERM, U738, Paris, France
[2]Université Paris 7 Denis Diderot, UFR de Médecine, Paris
[3]AP-HP, Hôpital Bichat, Département d'Epidémiologie, Biostatistique et Recherche Clinique, Paris
[4]INSERM Centre d'Investigation Clinique 202, Tours, France
[5]Université François Rabelais, Tours
[6]CHRU de Tours, Tours

Correspondence to: P Ravaud, Département d'Epidémiologie, Biostatistique et Recherche Clinique, Secteur Claude Bernard, Hôpital Bichat Claude Bernard, 75877 Paris, cedex 18, France
philippe.ravaud@bch.aphp.fr

Cite this as: *BMJ* 2009;338:b1732
doi:10.1136/bmj.b1732

ABSTRACT

Objectives To assess quality of reporting of sample size calculation, ascertain accuracy of calculations, and determine the relevance of assumptions made when calculating sample size in randomised controlled trials.

Design Review.

Data sources We searched MEDLINE for all primary reports of two arm parallel group randomised controlled trials of superiority with a single primary outcome published in six high impact factor general medical journals between 1 January 2005 and 31 December 2006. All extra material related to design of trials (other articles, online material, online trial registration) was systematically assessed. Data extracted by use of a standardised form included parameters required for sample size calculation and corresponding data reported in results sections of articles. We checked completeness of reporting of the sample size calculation, systematically replicated the sample size calculation to assess its accuracy, then quantified discrepancies between a priori hypothesised parameters necessary for calculation and a posteriori estimates.

Results Of the 215 selected articles, 10 (5%) did not report any sample size calculation and 92 (43%) did not report all the required parameters. The difference between the sample size reported in the article and the replicated sample size calculation was greater than 10% in 47 (30%) of the 157 reports that gave enough data to recalculate the sample size. The difference between the assumptions for the control group and the observed data was greater than 30% in 31% (n=45) of articles and greater than 50% in 17% (n=24). Only 73 trials (34%) reported all data required to calculate the sample size, had an accurate calculation, and used accurate assumptions for the control group.

Conclusions Sample size calculation is still inadequately reported, often erroneous, and based on assumptions that are frequently inaccurate. Such a situation raises questions about how sample size is calculated in randomised controlled trials.

INTRODUCTION

The importance of sample size determination in randomised controlled trials has been widely asserted, and according to the CONSORT statement these calculations must be reported and justified in published articles.[1-4] The aim of an a priori sample size calculation is mainly to determinate the number of participants needed to detect a clinically relevant treatment effect.[5,6] Some have asserted that oversized trials, which expose too many people to the new therapy, or underpowered trials, which may fail to achieve significant results, should be avoided.[7-12]

The usual conventional approach is to calculate sample size with four parameters: type I error, power, assumptions in the control group (response rate and standard deviation), and expected treatment effect.[5] Type I error and power are usually fixed at conventional levels (5% for type I error, 80% or 90% for power). Assumptions related to the control group are often pre-specified on the basis of previously observed data or published results, and the expected treatment effect is expected to be hypothesised as a clinically meaningful effect. The uncertainty related to the rate of events or the standard deviation in the control group[13,14] and to treatment effect could lead to lower than intended power.[6]

We aimed to assess the quality of reporting sample size calculation in published reports of randomised controlled trials, the accuracy of the calculations, and the accuracy of the a priori assumptions.

MATERIALS AND METHODS

Search strategy

We searched MEDLINE via PubMed with the search terms "randomized controlled trials" and "randomised controlled trials" for articles published in six general journals with high impact factors: *New England Journal of Medicine*, *Journal of the American Medical Association* (*JAMA*), *The Lancet*, *Annals of Internal Medicine*, *BMJ*, and *PLoS Medicine* between 1 January 2005 and 31 December 2006. One of us (PC) screened titles and abstracts of retrieved articles to identify relevant articles, with the help of a second reviewer (AD) if needed.

Selection of relevant articles

We included all two arm, parallel group superiority randomised controlled trials with a single primary outcome. We excluded reports for which the study design

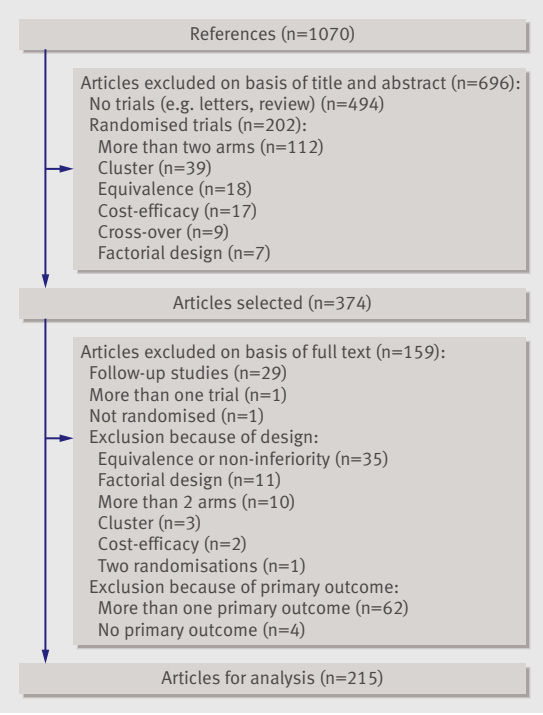

Fig 1 | **Study screening process**

was factorial, cluster, or crossover. We selected the first report that presented the results for the primary outcome. We excluded follow-up studies.

Data abstraction

For all selected articles, we systematically retrieved and assessed the full published report of the trial, any extra material or appendices available online, the study design article, if cited, and the details of online registration of the trial, if mentioned. A standardised data collection form was generated on the basis of a review of the literature and a priori discussion and tested by the research team. We recorded the following data.

In the full text of the articles

General characteristics of the studies: including the medical area, whether the trial was multicentre, the type of treatment (pharmacological, non-pharmacological, or both), the type of primary endpoint (dichotomous, time to event, continuous), and the funding source (public, private, or both).

Details of the a priori sample size calculation as reported in the materials and methods section: we noted whether the sample size calculation was reported and, if so, the target sample size. We also collected all the parameters used for the calculation: type I error, one or two tailed test, type II error or power, type of test, assumptions in the control group (rate of events for dichotomous and time to event outcomes and standard deviation for continuous outcomes), and the predicted treatment effect (rate of events in the treatment group for dichotomous and time to event outcomes, mean difference or effect size [defined in appendix 1] for continuous outcomes). Any justification for assumptions made was also recorded.

Observed data as reported in the results section: number of patients randomised and analysed was recorded, and results for the control group. We also noted whether the results of the trial were statistically significant for the primary outcome.

In the online extra material or study design article

We recorded the target sample size and all the required parameters for sample size calculation if different from those reported in the article.

In the trial registration website

We noted the target sample size and all the required parameters for sample size calculation.

One of us (PC) independently completed all data extractions. A second member of the team (AD) reviewed a random sample of 30 articles for quality assurance. The κ statistic provided a measure of inter-observer agreement. The reviewers were not blinded to the journal name and authors.

Data analysis

Replication of sample size calculation

We replicated the sample size calculation for each article that provided all the data needed for the calculation. If parameters for replicating the sample size were missing in the article and if the calculation was described elsewhere (in the online extra material or study design article) we used the parameters given in this supplemental material. If the missing values were only the α risk or whether the test was one or two tailed, we hypothesised an α risk of 0.05 with a two tailed test to replicate the calculation. Sample size calculations were replicated by one of us (PC) with nQuery Advisor version 4.0 (Statistical Solutions, Cork, Ireland). For a binary endpoint, the replication used the formulae adapted for a χ^2 test or Fisher's exact test if specified in the available data. For a time to event endpoint the replication used the formulae adapted for a log rank test, and for a continuous endpoint the replication used the formulae adapted for Student's *t* test. The formulae used for the replication are provided and explained in appendix 1. If the absolute value of the standardised difference between the recalculated sample size and the reported sample size was greater than 10%, an independent statistician (GB) extracted the data from the full text independently and replicated the sample size calculation again. Any difference between the two calculations was resolved by consensus. The standardised difference between the reported sample size calculation and the replicated one is defined by the reported sample size calculation minus the recalculated sample size divided by the reported sample size calculation.

Comparisons between a priori assumptions and observed data

To assess the accuracy of a priori assumptions, we calculated relative differences between hypothesised

parameters for the control group reported in the materials and methods sections of articles and estimated ones reported in the results sections. We calculated relative differences for standard deviations if the outcome was continuous (standard deviation in the materials and methods section minus standard deviation in the results section divided by standard deviation in the materials and methods section) or for event rates for a dichotomous or time to event outcome (event rates in the materials and methods section minus event rates in the results section divided by event rates in the materials and methods section). The relation between the size of the trial and the difference between the assumptions and observed data was explored by use of Spearman's correlation coefficient, and its 95% confidence interval was estimated by bootstrap.

Statistical analyses were done with SAS version 9.1 (SAS Institute, Cary, NC), and R version 4.1 (Free Software Foundation's GNU General Public License).

RESULTS

Selected articles

Figure 1 summarises selection of articles. The electronic search yielded 1070 citations, including 281 reports of parallel group superiority randomised controlled trials with two arms. We selected 215 articles (appendix 2) that reported only one primary outcome.

Description of trials

Table 1 describes the characteristics of the included studies. The median sample size of the trials was 425 (interquartile range [IQR] 158-1041), and 112 reports (52.1%) claimed significant results for the primary endpoint. Seventy-six percent were multicentre trials, with a median of 23 centres (IQR 7-59). The three most frequent medical areas of investigation were cardiovascular diseases (26%; n=56 articles), infectious diseases (11%; n=24), and haematology and oncology (10%; n=22). Interobserver agreement in extracting the data from reports was good; κ coefficients ranged from 0.76 to 1.00.

Table 1 | Characteristics of 215 included studies

Characteristic	N (%)
Journal of publication	
New Engl J Med	80 (37)
Lancet	46 (21)
JAMA	39 (18)
BMJ	36 (17)
Ann Intern Med	12 (6)
PLoS Med	2 (1)
Year of publication	
2005	109 (51)
2006	106 (49)
Median number of randomised patients (IQR)	425 (158-1041)
Multicentre trial	163 (76)
Intervention	
Pharmacological	131 (61)
Nonpharmacological	74 (34)
Both	10 (5)
Outcome	
Dichotomous	100 (47)
Time to event	67 (31)
Continuous	48 (22)
Funding	
Only public	126 (59)
Only private	49 (23)
Both public and private	38 (18)
Unclear	2 (1)

IQR=interquartile range.

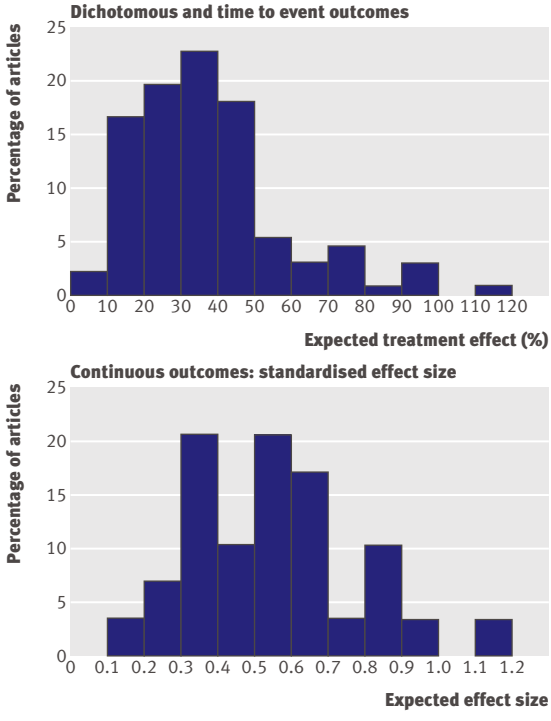

Fig 2 | Histogram of assumptions of treatment effect. For dichotomous and time to event outcomes: relative difference of event rates (larger rate minus smaller rate, divided by rate in control group). For continuous outcomes: standardised effect size.

Reporting of required parameters for a priori sample size calculation

Ten articles (5%) did not report any sample size calculation. Only 113 (53%) reported all the required parameters for the calculation. Table 2 describes the reporting of necessary parameters for sample size calculation.

The median of the expected treatment effect for dichotomous or time to event outcomes (relative difference of event rates) was 33.3% (IQR 24.8-50.0) and the median of the expected effect size for continuous outcomes was 0.53 (0.40-0.69) (fig 2).

The design of 35 of the 215 trials (16%) was described elsewhere. In two, the primary outcome described in the report differed from that in the design article. In 31 articles (89%), the data for sample size calculation were given. For 16 articles (52%) the reporting of the assumptions differed from the design article.

Table 2 | Reporting of parameters required for a priori sample size calculation for the 215 articles

Parameter	Reporting frequency (%)
α risk	191 (93)
0.05	183 (96)
Two tailed test	119 (65)
One tailed test	7 (4)
Unspecified	57 (31)
0.025 for one tailed test	2 (1)
Adapted for interim analyses	6 (3)
Power	200 (98)
80%	107 (54)
85%	9 (5)
90%	66 (33)
95%	4 (2)
Other values	14 (7)
Assumptions for control group	165 (81)
Justification of assumptions	81 (49)
Results from previous trial	54 (67)
Preliminary study	15 (19)
Observational data	6 (7)
Results of systematic review	2 (3)
Others	4 (5)
Assumptions for the treatment effect	186 (91)
Justification of the assumptions	50 (27)
Analogy to another trial or treatment	41 (82)
Clinical relevance	7 (14)
Observational data	1 (2)
Results of a meta-analysis	1 (2)
All parameters required for sample size calculation	113 (53)

Reporting of sample size calculation in online trial registration database

Of the 215 selected articles, 113 (53%) reported registration of the trial in an online database. Among them, 87 (77%) were registered in ClinicalTrials.gov, 23 (20%) in controlled-trials.com (ISRCTN registry), and three (3%) in another database. For 96 articles (85%), an expected sample size was given in the online database and was equal to the target sample size reported in the article in 46 of these articles (48%). The relative difference between the registered and reported sample size was greater than 10% in 18 articles (19%) and greater than 20% in five articles (5%). The parameters for the sample size calculation were not stated in the online registration databases for any of the trials.

Replication of sample size calculation

We were able to replicate sample size calculations for 164 articles: 113 reported all the required parameters, and 51 that omitted only the α risk or whether the test was one or two tailed. We were able to compare our recalculated sample size and the target sample size for 157 articles, since seven did not report any target sample size. The sample size recalculation was equal to the authors' target sample size for 27 articles (17%) and close (absolute value of the difference <5%) for 76 (48%). The absolute value of the difference between the replicated sample size calculation and the authors' target sample size was greater than 10% for 47 articles (30%) and greater than 50% for 10 (6%). Twenty-eight recalculations (18%) were 10% lower than reported sample size, and 19 recalculations (12%) were larger than reported sample size (fig 3). The results were similar when we analysed only the 113 articles reporting all the required parameters.

Comparisons between a priori parameters and corresponding estimates in results section

A comparison between the a priori assumptions and observed data was feasible for 145 of the 157 articles reporting enough parameters to recalculate the sample size and reporting the results of the authors' calculations.

Assumptions about control group

The median relative difference between the control group pre-specified parameters and their estimates was 3.3% (IQR −16.7 to 21.4). The median difference was 2.0% (−15 to 21) for dichotomous or time to event outcomes and 11% (−24 to 27) for continuous outcomes. The absolute value of the relative difference was greater than 30% for 45 articles (31%) and greater than 50% for 24 (17%). Figure 4 shows that the differences between the assumptions and the results were large and small in roughly even proportions, whether the results were significant or not. The size of the trial and the differences between the assumptions for the control group and the results did not seem to be substantially related (rho=0.03, 95% confidence interval −0.05 to 0.15).

Overall, 73 articles (34%) reported enough parameters for us to replicate the sample size calculation, had an accurate calculation (the replicated sample size calculation differed by less than 10% from the reported target sample size), and had accurate assumptions for the control group (the differences between the a priori assumptions and their estimates was less than 30%) (fig 5).

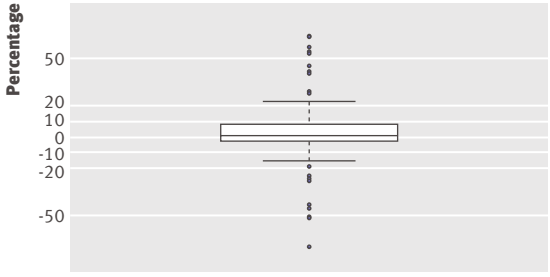

Fig 3 | Differences between target sample size and replicated sample size calculations. Differences in sample size calculations are relative differences between target sample size given in materials and methods section of articles and our recalculation with the parameters provided. Box represents median observations (horizontal rule) with 25th and 75th percentiles of observed data (top and bottom of box). Length of each whisker is 1.5 times interquartile range.

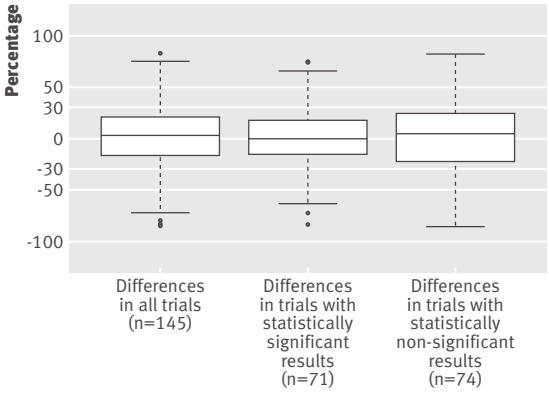

Fig 4 | Relative differences between assumptions and results for control groups

DISCUSSION

Principal findings

In this survey of 215 reports published in 2005 and 2006 in six general medical journals with high impact factors, only about a third (n=73, 34%) adequately described sample size calculations—ie, they reported enough data to recalculate the sample size, the sample size calculation was accurate, and assumptions in the control group differed less than 30% from observed data. Our study raises two main issues. The first is the inadequate reporting and the errors in sample size calculations, which are surprising in high quality journals with a peer review process; the second is the large discrepancies between the assumptions and the data in the results, which raises a much more complex problem because investigators often have to calculate a sample size with insufficient data to estimate these assumptions.

Reporting of the sample size calculation has greatly increased in the past decades, from 4% of reports describing a calculation in 1980 to 83% of reports in 2002.[15 16] However, our review highlights that some of the required parameters for sample size calculation are frequently absent in reports and that sample size miscalculations unfortunately occur in randomised controlled trials. We were not able to identify the reasons for such erroneous calculations, particularly the frequency of reported calculations that were greater than our recalculation. Surprisingly, such errors (sometimes large) were missed during the review process.

We also found large discrepancies between values for assumed parameters in the control group used for sample size calculations (ie, event rate or standard deviation in the control group) and estimated ones from observed data. Assumed values were fixed at a higher or lower level than corresponding data in the results sections in roughly even proportions, a finding different from the results of a previous study: Vickers showed that the sample standard deviation was greater than the pre-specified standard deviation for 80% of endpoints in randomised trials.[14]

Although the CONSORT group recommends reporting details of sample size determination to identify the primary outcome and as a sign of proper trial planning, our results suggest that researchers, reviewers, and editors do not take reporting of sample size determination seriously.[17] In this case, an effort should be made to increase transparency in sample size calculation or, if sample size calculation reporting is of little relevance in randomised controlled trials, perhaps it should be abandoned, as has been suggested by Bacchetti.[18]

Limitations

An important limitation of this study is that we could not directly assess whether assumptions had been manipulated to obtain feasible sample sizes because we used only published data. Assumptions can be first adapted when planning the trial, by retrofitting the assumption estimates to the available participants, also called "sample size samba" by Schulz and Grimes.[6] This situation is impossible to assess without attending the discussion between the investigators and statisticians. The sample size calculation can also be manipulated after the completion of the study, as Chan and coworkers have recently shown by comparing protocols to final articles.[19]

We included only two arm parallel group superiority randomised controlled trials with a single primary outcome, so we did not assess more complex sample size

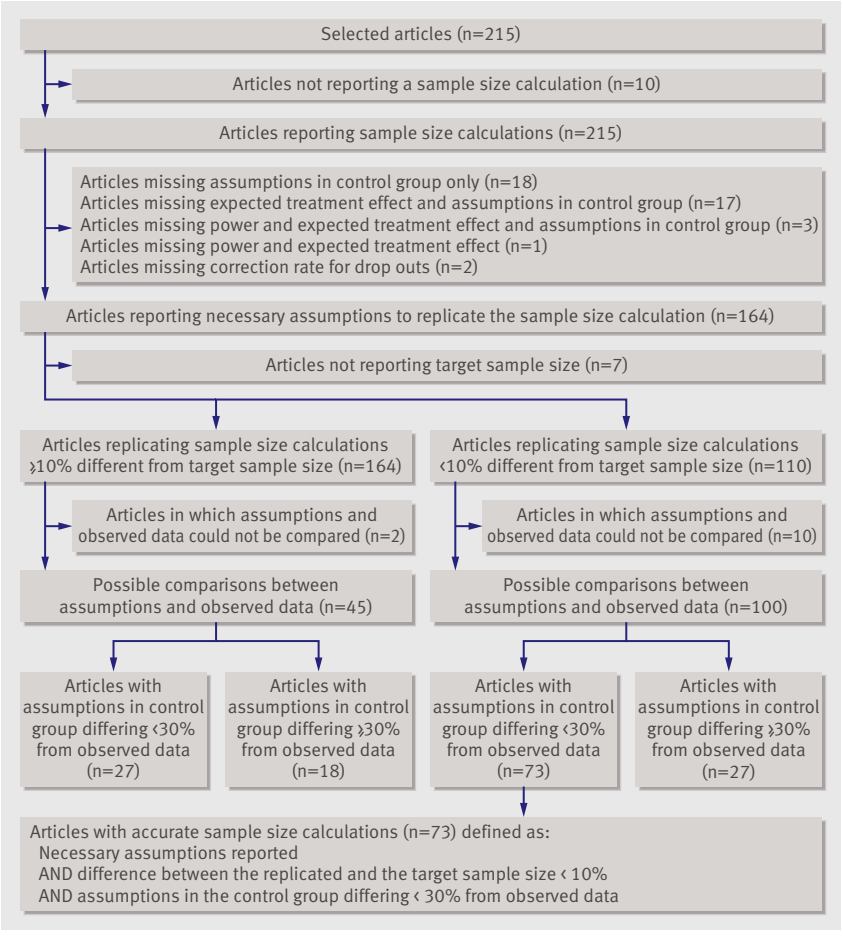

Fig 5 | Articles selected for analysis of sample size calculations

WHAT IS ALREADY KNOWN ON THIS SUBJECT

Planning and reporting of sample size calculation for randomised controlled trials is recommended by ICH E9 and the CONSORT statement

WHAT THIS STUDY ADDS

Sample size calculations are inadequately reported, often erroneous, and based on assumptions that are frequently inaccurate

These major issues question the foundation of sample size calculation and its reporting in randomised controlled trials

calculations. We chose these trials to give a homogeneous sample of articles. We also selected only general medical journals with a high impact factor. Low impact factor journals could have the same or lower methodological quality. We chose one hypothesis when we recalculated the sample size: the α risk was set at 0.05 for a two tailed test when one (or two) of these parameters was missing. Nevertheless, the proportion of inadequate calculations did not change whether we excluded these articles or not.

Implications

A major discrepancy exists between the importance given to sample size calculation by funding agencies, ethics review boards, journals, and investigators and the current practice of sample size calculation and reporting.[20] Sample size calculations are frequently based on inaccurate assumptions for the control group, calculations are often erroneous, and the hypothesised treatment effect is often fixed a posteriori.[6] This statement does not even take into account that the primary outcome reported in the initial protocol (on which the sample size calculation was theoretically based) was found to differ from the primary outcome of the final report in 62% of trials.[21] As written by Senn, "the sample size calculation is an excuse for a sample size, not a reason," and the current calculation of sample size is actually mainly driven by feasibility.[9 20]

We wonder whether the questions raised by our results should join the debate on the ethics of underpowered trials. Although underpowered trials are viewed as unethical by many people, others consider such trials ethical in that some evidence is better than none[6 22] and such trials could even produce more information than larger studies.[23] Furthermore, results of underpowered trials contribute to the body of knowledge and are useful for meta-analysis.[20] We therefore believe, as do others, that there is room for reflection on how sample size should be determined for randomised trials.[24 25] After years of trials with supposedly inadequate sample sizes, it is time to develop and use new ways of planning sample sizes.

Funding: PC received financial support from Direction Régionale des Affaires Sanitaires et Sociales. AD received financial support from Fédération Hospitalière de France and Assistance Publique des Hôpitaux de Paris. The authors were independent from the funding bodies.

Contributors: The guarantor (PR) had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: PC, PR. Acquisition of data: PC, AD. Recalculation of sample sizes: PC, GB. Analysis and interpretation of data: PC, PR, BG, AD, GB. Drafting of manuscript: PC, AD, PR. Critical revision of manuscript for important intellectual content: PR, BG, AD, GB. Statistical analysis: PC, PR, BG, GB, AD. Administrative, technical, or material support: PR. Study supervision: PR, BG.

Competing interests: None declared.

Ethical approval: Not required.

1 ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. International Conference on Harmonisation E9 Expert Working Group. *Stat Med* 1999;18:1905-42.
2 Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 1994;272:122-4.
3 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.
4 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191-4.
5 Machin D, Campbell M, Fayers P, Pinol A. *Sample size tables for clinical studies*. 2nd ed. Oxford: Blackwell Science, 1997.
6 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005;365:1348-53.
7 Lakatos E. Sample size determination. In: Redmond C, Colton T, eds. *Biostatistics in clinical trials*. Chichester: John Wiley and Sons, 2001.
8 Altman DG. Statistics and ethics in medical research: III How large a sample? *BMJ* 1980;281:1336-8.
9 Senn S. *Statistical issues in drug development*. Chichester: John Wiley and Sons, 1997.
10 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690-4.
11 Wooding WM. *Planning pharmaceutical clinical trials*. Chichester: John Wiley and Sons, 1994.
12 Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002;288:358-62.
13 Weaver CS, Leonardi-Bee J, Bath-Hextall FJ, Bath PM. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke* 2004;35:1216-24.
14 Vickers AJ. Underpowering in randomized trials reporting a sample size calculation. *J Clin Epidemiol* 2003;56:717-20.
15 Meinert CL, Tonascia S, Higgins K. Content of reports on clinical trials: a critical review. *Control Clin Trials* 1984;5:328-47.
16 Mills EJ, Wu P, Gagnier J, Devereaux PJ. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. *Contemp Clin Trials* 2005;26:480-7.
17 Altman DG, Moher D, Schulz KF. Peer review of statistics in medical research. Reporting power calculations is important. *BMJ* 2002;325:491.
18 Bacchetti P. Peer review of statistics in medical research: the other problem. *BMJ* 2002;324:1271-3.
19 Chan AW, Hrobjartsson A, Jorgensen KJ, Gotzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ* 2008;337:a2299.
20 Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter? *PLoS Med* 2008;5:e4.
21 Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457-65.
22 Edwards SJ, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. *Lancet* 1997;350:804-7.
23 Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. *Am J Epidemiol* 2005;161:105-10.
24 Simon R. Discussions. *Biometrics* 2008;64:589-91.
25 Bacchetti P, McCulloch CE, Segal MR. Simple, defensible sample sizes based on cost efficiency. *Biometrics* 2008;64:577-85; discussion 586-94.

Accepted: 2 January 2009