Epidemiology 1

Sample size calculations in randomised trials: mandatory and mystical

Lancet 2005; 365: 1348–53

Kenneth F Schulz, David A Grimes

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA
(K F Schulz PhD, D A Grimes MD)

Correspondence to:
Dr Kenneth F Schulz
KSchulz@fhi.org

Investigators should properly calculate sample sizes before the start of their randomised trials and adequately describe the details in their published report. In these a-priori calculations, determining the effect size to detect—eg, event rates in treatment and control groups—reflects inherently subjective clinical judgments. Furthermore, these judgments greatly affect sample size calculations. We question the branding of trials as unethical on the basis of an imprecise sample size calculation process. So-called underpowered trials might be acceptable if investigators use methodological rigor to eliminate bias, properly report to avoid misinterpretation, and always publish results to avert publication bias. Some shift of emphasis from a fixation on sample size to a focus on methodological quality would yield more trials with less bias. Unbiased trials with imprecise results trump no results at all. Clinicians and patients deserve guidance now.

Sample size calculations for randomised trials seem unassailable. Indeed, investigators should properly calculate sample sizes and adequately describe the key details in their published report. Research methodologists describe the approaches in books and articles. Protocol committees and ethics review boards require adherence. CONSORT reporting guidelines clearly specify the reporting of sample size calculations.[1,2] Almost everyone agrees.

An important impetus to this unanimity burst on the medical world more than a quarter of a century ago. A group of researchers, led by Tom Chalmers, published a landmark article detailing the lack of statistical power in so-called negative randomised trials published in premier general medical journals.[3] In Chalmers' long illustrious career, he published hundreds of articles. This article on sample size and power received many citations. Paradoxically, that troubled him.[4] He regarded it as the most damaging paper that he had ever coauthored. Why? We will describe his concerns later, so stay tuned.

Components of sample size calculations

Calculating sample sizes for trials with dichotomous outcomes (eg, sick *vs* well) requires four components: type I error (α), power, event rate in the control group, and a treatment effect of interest (or analogously an event rate in the treatment group). These basic components persist through calculations with other types of outcomes, except other assumptions can be necessary. For example, with quantitative outcomes and a typical statistical test, investigators might assume a difference between means and a variance for the means.

In clinical research, hypothesis testing risks two fundamental errors (panel 1). First, researchers can conclude that two treatments differ when, in fact, they do not. This type I error (α) measures the probability of making this false-positive conclusion. Conventionally, α is most frequently set at 0·05, meaning that investigators desire a less than 5% chance of making a false-positive conclusion. Second, researchers can conclude that two treatments do not differ when, in fact, they do—ie, a false-negative conclusion. This type II error (β) measures the probability of this false-negative conclusion. Conventionally, investigators set β at 0·20, meaning that they desire less than a 20% chance of making a false-negative conclusion.

Power derives from β error. Mathematically, it is the complement of β (1–β) and represents the probability of avoiding a false-negative conclusion. For example, for

β=0·20, the power would be 0·80, or 80%. Stated alternatively, power represents the likelihood of detecting a difference (as significant, with $p<\alpha$), assuming a difference of a given magnitude exists. For example, a trial with a power of 80% has an 80% chance of detecting a difference between two treatments if a real difference of assumed magnitude exists in the population.

Admittedly, understanding α error, β error, and power can be a challenge. Convention, however, usually guides investigators for inputs into sample size calculations. The other inputs cause lesser conceptual difficulties, but produce pragmatic problems. Investigators estimate the true event rates in the treatment and control groups as inputs. Usually, we recommend estimating the event rate in the population and then determining a treatment effect of interest. For example, investigators estimate an event rate of 10% in the controls. They then would estimate an absolute change (eg, an absolute reduction of 3%), a relative change (a relative reduction of 30%), or simply estimate a 7% event rate in the treatment group. Using these assumptions, investigators calculate sample sizes. Standard texts describe the procedures encompassing, for example, binary, continuous, and time-to-event measures.[5–7] Commonly, investigators use sample size and power software (preferably with guidance from a statistician). Most hand calculations diabolically strain human limits, even for the easiest formula, such as we offer in panel 2.

Effect of selecting α error and power

The conventions of α=0·05 and power=0·80 usually suffice. However, other assumptions make sense based on the topic studied. For example, if a standard prophylactic antibiotic for hysterectomy is effective with few side-effects, in a trial of a new antibiotic we might set α error lower (eg, 0·01) to reduce the chances of a false-positive conclusion. We might even consider lowering the power below 0·80 because of our reduced concern about missing an effective treatment—an effective safe treatment already exists. By contrast, if an investigator tests a standard prophylactic antibiotic against a cheap safe vitamin supplement the balance changes. Little harm could come from making an α error so setting it at 0·10 might make sense.[7] However, if this cheap easy intervention produced benefit, we would not want to miss it. Thus, investigators might increase power to 0·99.

Different assumptions about α error and power directly change sample sizes. Reducing α and increasing power both increase the sample: for example, reducing α from 0·05 to 0·01 generates about a 70% increase in trial size at power=0·50 and a 50% increase at power=0·80 (table). At α=0·05, increasing power from 0·50 to 0·80 yields a two-fold increase in trial size and from 0·50 to 0·99 almost a five-fold increase (table). Choices of α and power thus produce different sample sizes and trial costs.

Panel 1: Errors defined

Type I error (α)
The probability of detecting a statistically significant difference when the treatments are in reality equally effective—ie, the chance of a false-positive result.

Type II error (β)
The probability of not detecting a statistically significant difference when a difference of a given magnitude in reality exists—ie, the chance of a false-negative result.

Power (1–β)
The probability of detecting a statistically significant difference when a difference of a given magnitude really exists.

Panel 2: The simplest, approximate sample size formula for binary outcomes, assuming α=0·05, power=0·90, and equal sample sizes in the two groups

n=the sample size in each of the groups
p_1=event rate in the treatment group (not in formula but implied when R and p_2 are estimated)
p_2=event rate in the control group
R=risk ratio (p_1/p_2)

$$n=\frac{10{\cdot}51[(R+1)-p_2(R^2+1)]}{p_2(1-R)^2}$$

For example, we estimate a 10% event rate in the control group (p_2=0·10) and determine that the clinically important difference to detect is a 40% reduction (R=0·60) with the new treatment at α=0·05 and power=0·90. (Note: R=0·60 equates to an event rate in the treatment group of p_1=0·06, ie, R=6%/10%)

$$n=\frac{10{\cdot}51[(0{\cdot}60+1)-0{\cdot}10\,(0{\cdot}60^2+1)]}{0{\cdot}10\,(1-0{\cdot}60)^2}$$

n=961·665≈962 in each group (PASS software version 6.0 [NCSS, Kaysville, UT, USA] with a more accurate formula yields 965)

This formula accommodates alternate α levels and power by replacing 10·51 with the appropriate value from the table below.

	Power (1–β)		
	0·80	0·90	0·95
α (type I error)			
0·05	7·85	10·51	13·00
0·01	11·68	14·88	17·82

	Power (1–β)			
	0·50	0·80	0·90	0·99
α (type I error)				
0·05	100	200	270	480
0·01	170	300	390	630
0·001	280	440	540	820

Table: **Approximate relative trial sizes for different levels of α and power**

Some investigators use one-sided tests for α error to reduce estimated sample sizes. We discourage that approach. While we have assumed two-sided tests thus far, one-sided tests might indeed make sense in view of available biological knowledge. However, that decision should not affect sample size estimation. We suggest the same standard of evidence irrespective of whether a one-sided or two-sided test is assumed.[7] Thus, a one-sided $\alpha=0\cdot025$ yields the same level of evidence as a two-sided $\alpha=0\cdot05$. Using a one-sided test in sample size calculations to reduce required sample sizes stretches credulity.

Estimation of population parameters

For some investigators, estimation of population parameters—eg, event rates in the treatment and control groups—has mystical overtones. Some researchers scoff at this notion, since estimating the parameters is the aim of the trial: needing to do it before the trial seems ludicrous. The key point, however, is that they are not estimating the population parameters per se but the treatment effect they deem worthy of detecting. That is a big difference.

Usually, investigators start by estimating the event rate in the control group. Sometimes scant data lead to unreliable estimates. For example, we needed to estimate an event rate for pelvic inflammatory disease in users of intrauterine devices in a family planning population in Nairobi, Kenya. Government officials estimated 40%; the clinicians at the medical centre thought that estimate was much too high and instead suggested 12%. We conservatively planned on 6%, but the placebo group in the actual randomised trial yielded 1·9%.[8] The first estimate was off by more than 20-fold, which enormously affects sample size calculations.

Published reports can provide an estimate of the endpoint in the control group. Usually, however, they incorporate a host of differences, such as dissimilar locations, eligibility criteria, endpoints, and treatments. Nevertheless, some information on the control group usually exists. That becomes the starting point.

In a trial on prevention of fever after hysterectomy, data assumed to be reasonably good show that 10% of women have febrile morbidity after the standard prophylactic antibiotic. That becomes the event rate for the control group. Estimation of the effect size of interest should reflect both clinical acumen and the potential public-health effect. This important aspect should not default to a statistician. The decision process proceeds by accumulating clinical background knowledge. Assume the standard antibiotic costs US$10 for prophylaxis, incurs few side-effects, and is administered orally. The new antibiotic costs US$200 for prophylaxis, has more side-effects, is administered intravenously, but has a broader range of coverage. All these pragmatic and clinical factors bear on the decision process. In view of the 10% event rate for fever in the control group, and knowing the clinical background, would we be interested in detecting a 10% reduction to 9%; a 20% reduction to 8%; a 30% reduction to 7%; a 40% reduction to 6%; a 50% reduction to 5%; and so forth? Determining the difference to detect reflects inherently subjective clinical judgments. No right answer exists. We could say that a 30% reduction is worthwhile to detect, but another investigator might decide on a 50% reduction.

These parameter assumptions enormously affect sample size calculations. Keeping the assumptions for the control group constant, halving the effect size necessitates a greater than four-fold increase in trial size. Similarly, quartering the effect size requires a greater than 16-fold increase in trial size. Stated alternatively, sample sizes rise by the inverse square of the effect size reduction (which statisticians call a quadratic relation). For example, in view of our initial parameter estimates of 10% in the control group and 6% in the intervention group, and $\alpha=0\cdot05$ and power=0·90, about 965 participants would be necessary in each group (panel 2). Halving the effect size, thereby altering the intervention group estimate to 8%, requires a more than four-fold increase in sample size to 4301. Quartering the effect size, thereby altering the intervention group estimate to 9%, necessitates a more than 18-fold increase in trial size to 18 066 per group. Small changes in effect size generate large changes in trial size.

The need for huge trial sizes with low event rates frustrates investigators. That frustration partly stems from a lack of understanding that, with binary endpoints, numerator events drive trial power rather than denominators. For example, assume $\alpha=0\cdot05$ and a desired 40% reduction in the outcome event rate. A trial of 2000 participants (1000 assigned to the treatment group and 1000 to the control) with a control group event rate of 10% would provide similar power to a trial of 20 000 participants (10 000 assigned to each group) with a control group event rate of 1%. Both trials would need a similar number of numerator events—about 160—for roughly 90% power.

Low power with limited available participants

What happens when sample size software—in view of an investigator's diligent estimates—yields a trial size that exceeds the number of available participants? Frequently, investigators then calculate backwards and estimate that they have low power (eg, 0·40) for their

available participants. This practice may be more the rule than the exception.[9]

Some methodologists advise clinicians to abandon such a low-power study. Many ethics review boards deem a low power trial unethical.[10–12] Chalmers' early paper on the lack of power in published trials contributed to this response, which brings us back to our opening paragraphs. He felt his group's article fuelled these over-reactions.[4]

Chalmers eventually stated that so-called underpowered trials can be acceptable because they could ultimately be combined in a meta-analysis.[4,13] This view seems unsupported by many statisticians, surprisingly even those in favour of small trials.[9] Nevertheless, we agree with Chalmers' view, which undoubtedly will draw the ire of many statisticians and ethicists. Our support attaches three caveats.

First, the trial should be methodologically strong, thus eliminating bias. Unfortunately, the adequate-power mantra frequently overwhelms discussion on other methodological aspects. For example, inadequate randomisation usually yields biased results. Those biased results cannot be salvaged even if a huge sample size generates great precision.[14–16] By contrast, if investigators design and implement a trial properly, that trial essentially yields an unbiased estimate of effect, even if it has lower power (and precision). Moreover, because the results are unbiased, the trial could be combined with similar unbiased trials in a meta-analysis. Indeed, this idea, especially when incorporated into prospective meta-analyses,[17] is akin to multicentre trials.

Second, authors must report their methods and results properly to avoid misinterpretation. If they report the trial results properly using interval estimation, the wide confidence intervals around the estimated treatment effect would accurately depict the low power. Reporting of confidence intervals represents a worthwhile contribution and avoids "the absence of evidence is not evidence of absence" problem wrought by simplistic $p>0\cdot05$ conclusions.[18–20]

Third, low-powered trials must be published irrespective of their results, thereby becoming available for meta-analysis. Publication bias constitutes the strongest argument against underpowered trials.[21,22] Publication bias emerges when published trials do not represent all trials undertaken, usually because statistically significant results tend to be submitted and published more frequently than indeterminate results. Low-powered trials contribute to the problem because they more generally yield an indeterminate result. Condemnation of all underpowered trials and prevention of their conduct, however, thwarts important research. We need to directly tackle the real culprit of publication bias, and the scientific community has made great strides. Not publishing completed trials is called both unscientific and unethical in the scientific literature.[23–25] Trial registration schemes catalogue ongoing trials such that their results will not be lost. Furthermore, large systematic review enterprises, most notably the Cochrane Collaboration, scour unpublished work to reduce publication bias.

Proclamations of underpowered trials being unethical strike us as a bit odd for at least two reasons. First, preoccupation with sample size overshadows the more pertinent concerns of elimination of bias. Second, how can a process rife with subjectivity fuel a black-white decision on its ethics? With that subjectivity, basing trial ethics on statistical power seems simplistic and misplaced. Indeed, since investigators estimate sample size on the basis of rough guesses, if deeming the implementation of low power trials as unethical is taken to a logical extreme, then the world will have no trials because sample size determination would always be open to question. "Statements that it is unethical to embark on controlled trials unless an arbitrarily defined level of statistical power can be assured make no sense if the alternative is acquiescence in ignorance of the effects of healthcare interventions."[24] Edicts that underpowered trials are unethical challenge reason and, furthermore, disregard that sometimes potential participants desire involvement in trials.[26]

Sample size samba

Investigators sometimes perform a "sample size samba" to achieve adequate power.[27,28] The dance involves retrofitting of the parameter estimates (in particular, the treatment effect worthy of detection) to the available participants. This practice seems fairly common in our experience and in that of others.[27] Moreover, funding agencies, protocol committees, and even ethics review boards might encourage this backward process. It represents an operational solution to a real problem. In view of the circumstances, we do not judge harshly the samba, because it probably has facilitated the conduct of many important studies. Moreover, it truly depicts estimates of the sample sizes necessary given the provided assumptions. Nevertheless, the process emphasises the inconsistencies in the "underpowered trials are unethical" argument: a proposed trial is unethical before the "samba" and becomes ethical thereafter simply by shifting the estimate of effect size. All trials have an infinite number of powers, and low power is relative.

Sample size modification

With additional available participants and resource flexibility, investigators could consider a sample size modification strategy, which would alleviate some of the difficulties with rough guesses used in the initial sample size calculations. Usually, modifications lead to increased sample sizes,[29] so investigators should have access to the participants and the funding to accommodate the modifications.

Approaches to modification rely on revision of the event rate, the variance of the endpoint, or the treatment effect.[30–33] Importantly, any sample size modifications at an interim stage of a trial should hinge on a prespecified plan that avoids bias. The sponsor or steering committee should describe in the protocol a comprehensible plan for the timing and method of the potential modifications.[31]

Futility of post hoc power calculations

A trial yields a treatment effect and confidence interval for the results. The power of the trial is expressed in that confidence interval. Hence, the power is no longer a meaningful concern.[7,27,34] Nevertheless, after trial completion, some investigators do power calculations on statistically non-significant trials using the observed results for the parameter estimates. This exercise has specious appeal, but tautologically yields an answer of low power.[7,27] In other words, this ill-advised exercise answers an already answered question.

What should readers look for in sample size calculations?

Readers should find the a-priori estimates of sample size. Indeed, in trial reports, confidence intervals appropriately indicate the power. However, sample size calculations still provide important information. First, they specify the primary endpoint, which safeguards against changing outcomes and claiming a large effect on an outcome not planned as the primary outcome.[35] Second, knowing the planned size alerts readers to potential problems. Did the trial encounter recruitment difficulties? Did the trial stop early because of a statistically significant result? If so, the authors should provide a formal statistical stopping rule.[36] If they did not use a formal rule, then multiple looks at the data inflated α.[5,29] Similar problems can be manifested in larger than planned sample sizes. Providing planned sizes, however arbitrary, lays the groundwork for transparent reporting.

Low reported power or unreported sample size calculations usually are not a fatal flaw. Low power can reflect a lack of methodological knowledge, but it may just indicate an inadequate number of potential participants. Sample size calculations, even with low power, still provide the vital information described above. What if authors neglect mentioning a-priori sample size calculations? Readers should cautiously interpret the results because of the missing information on primary outcome and on stopping clues. Moreover, neglecting to report sample size calculations suggests a methodological naiveté that might portend other problems.

Nevertheless, readers should be most concerned with systematic errors (bias) hidden by investigators. Authors failing to report poor randomisation, inadequate allocation concealment, deficient blinding, or defective participant retention hide inadequacies that could cause major bias.[37–41] Thus, readers should ascribe less concern to perceived inadequate sample size for two substantial reasons: first, it does not cause bias and, second, any random error produced transparently surfaces in the confidence intervals and p values. The severest problems for readers are the systematic errors that are not revealed. In other words, readers should not totally discount a trial simply because of low power, but they should carefully weigh its value accordingly. The value resides in the context of other research, either past or future.[42]

Readers should find all assumptions underlying any sample size calculation: type I error (α), power (or β), event rate in the control group, and a treatment effect of interest (or analogously, an event rate in the treatment group). A statement that "we calculated necessary sample sizes of 120 in each group at α=0·05 and power=0·90" is almost meaningless, because it neglects the estimates for the effect size and control group event rate. Even small trials have high power to detect huge treatment effects.

Readers should also examine the assumptions for the sample size calculation. For example, they might believe that a smaller effect size is more worthy than the planned effect size. Therefore, the reader would be aware of the lower power of the trial relative to their preferred effect size.

Conclusions

Statistical power is an important notion, but it should be stripped of its ethical bellwether status. We question the branding of trials as unethical based solely on an inherently subjective, imprecise sample size calculation process. We endorse planning for adequate power, and we salute large multicentre trials of the ISIS-2 ilk;[43] indeed, more such studies should be undertaken. However, if the scientific world insisted solely on large trials, many unanswered questions in medicine would languish unanswered. Some shift of emphasis from a fixation on sample size to a focus on methodological quality would yield more trials with less bias. Unbiased trials with imprecise results trump no results at all.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

We thank David L Sackett, Douglas G Altman, Willard Cates, and Sir Iain Chalmers for their helpful comments on an earlier version of this manuscript.

References

1 Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports or parallel-group trials. *Lancet* 2001; **357:** 1191–94.

2 Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; **134:** 663–94.

3 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:** 690–94.

4 Sackett DL, Cook DJ. Can we learn anything from small trials? *Ann N Y Acad Sci* 1993; **703:** 25–31.

5 Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.

6 Meinert CL. Clinical trials: design, conduct, and analysis. New York: Oxford University Press, 1986.

7 Piantadosi S. Clinical trials: a methodologic perspective. New York: John Wiley and Sons, 1997.

8 Sinei SK, Schulz KF, Lamptey PR, et al. Preventing IUCD-related pelvic infection: the efficacy of prophylactic doxycycline at insertion. *Br J Obstet Gynaecol* 1990; **97:** 412–19.

9 Matthews JN. Small clinical trials: are they all bad? *Stat Med* 1995; **14:** 115–26.

10 Edwards SJ, Lilford RJ, Braunholtz D, Jackson J. Why "underpowered" trials are not necessarily unethical. *Lancet* 1997; **350:** 804–07.

11 Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA* 2002; **288:** 358–62.

12 Lilford RJ. The ethics of underpowered clinical trials. *JAMA* 2002; **288:** 2118–19.

13 Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline, I: control of bias and comparison with large co-operative trials. *Stat Med* 1987; **6:** 315–28.

14 Peto R. Failure of randomisation by "sealed" envelope. *Lancet* 1999; **354:** 73.

15 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273:** 408–12.

16 Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995; **274:** 1456–58.

17 Walker MD. Atrial fibrillation and antithrombotic prophylaxis: a prospective meta-analysis. *Lancet* 1989; **1:** 325–26.

18 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995; **311:** 485.

19 Detsky AS, Sackett DL. When was a "negative" clinical trial big enough? How many patients you needed depends on what you found. *Arch Intern Med* 1985; **145:** 709–12.

20 Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases: a way out of a conundrum. *BMJ* 1995; **311:** 1621–25.

21 Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Educ Prev* 1997; **9:** 15–21.

22 Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H Jr. Publication bias and clinical trials. *Control Clin Trials* 1987; **8:** 343–53.

23 Chalmers I. Underreporting research is scientific misconduct. *JAMA* 1990; **263:** 1405–08.

24 Chalmers I. Cardiotocography v Doppler auscultation: all unbiased comparative studies should be published. *BMJ* 2002; **324:** 483–85.

25 Antes G, Chalmers I. Under-reporting of clinical trials is unethical. *Lancet* 2003; **361:** 978–79.

26 Chalmers I. What do I want from health research and researchers when I am a patient? *BMJ* 1995; **310:** 1315–18.

27 Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994; **121:** 200–06.

28 Peipert JF, Metheny WP, Schulz K. Sample size and statistical power in reproductive research. *Obstet Gynecol* 1995; **86:** 302–05.

29 Ellenberg SS, Fleming TR, DeMets DL. Data monitoring committees in clinical trials. Chichester: John Wiley and Sons, 2002.

30 Wang SJ, Hung HM, Tsong Y, Cui L. Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. *Stat Med* 2001; **20:** 1903–12.

31 Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55:** 853–57.

32 Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55:** 1286–90.

33 Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med* 1990; **9:** 65–71.

34 Fayers PM, Machin D. Sample size: how many patients are necessary? *Br J Cancer* 1995; **72:** 1–9.

35 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.

36 Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet* (in press).

37 Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. *Lancet* 2002; **359:** 781–85.

38 Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet* 2002; **359:** 696–700.

39 Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. *Lancet* 2002; **359:** 614–18.

40 Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002; **359:** 515–19.

41 Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. *Lancet* 2002; **359:** 966–70.

42 Clarke M, Alderson P, Chalmers I. Discussion sections in reports of controlled trials published in general medical journals. *JAMA* 2002; **287:** 2799–801.

43 ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988; **2:** 349–60.