Subgroup analysis and other (mis)uses of baseline data in clinical trials

Susan F Assmann, Stuart J Pocock, Laura E Enos, Linda E Kasten

Summary

Background Baseline data collected on each patient at randomisation in controlled clinical trials can be used to describe the population of patients, to assess comparability of treatment groups, to achieve balanced randomisation, to adjust treatment comparisons for prognostic factors, and to undertake subgroup analyses. We assessed the extent and quality of such practices in major clinical trial reports.

Methods A sample of 50 consecutive clinical-trial reports was obtained from four major medical journals during July to September, 1997. We tabulated the detailed information on uses of baseline data by use of a standard form.

Findings Most trials presented baseline comparability in a table. These tables were often unduly large, and about half the trials inappropriately used significance tests for baseline comparison. Methods of randomisation, including possible stratification, were often poorly described. There was little consistency over whether to use covariate adjustment and the criteria for selecting baseline factors for which to adjust were often unclear. Most trials emphasised the simple unadjusted results and covariate adjustment usually made negligible difference. Two-thirds of the reports presented subgroup findings, but mostly without appropriate statistical tests for interaction. Many reports put too much emphasis on subgroup analyses that commonly lacked statistical power.

Interpretation Clinical trials need a predefined statistical analysis plan for uses of baseline data, especially covariate-adjusted analyses and subgroup analyses. Investigators and journals need to adopt improved standards of statistical reporting, and exercise caution when drawing conclusions from subgroup findings.

Lancet 2000; **355:** 1064–69

New England Research Institutes, Watertown, MA, USA (S F Assmann PhD, L E Enos MSc, L E Kasten MA)**; and Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK** (Prof S J Pocock PhD)

Correspondence to: Prof Stuart J Pocock

Introduction

For most randomised clinical trials, substantial baseline data are collected on each patient at randomisation. These data relate to demographics, medical history, current signs and symptoms, and quantitative disease measures (including some measured again later in the study as outcomes). Gathering of such baseline data seems to have four main aims. First, baseline data are used to characterise the patients included in the trial, and to show that the treatment groups are well balanced. Second, randomisation may include some means of balancing or stratification on a few key factors. Third, for analysis of outcome by treatment group, covariate adjustment may be used to take account of certain baseline factors. Fourth, subgroup analyses may be carried out to assess whether treatment differences in outcome (or lack thereof) depend on certain characteristics of patients.

Statistical reporting of clinical trials has improved, many journals have statistical refereeing, and clearer guidelines to authors[1] may further improve reporting quality. However, insufficient attention is paid to the quality and extent of reporting on these uses of baseline data.

We aimed to describe and critically evaluate current practice on the use of baseline data in clinical-trial reports in major medical journals, and to make recommendations to enhance the quality of future reporting, especially on the dangers of overemphasising subgroup analyses.

Methods

We handsearched all reports of clinical trials with individual randomisation of patients during July to September, 1997, in *BMJ*, *JAMA*, *The Lancet*, and *New England Journal of Medicine*. Crossover trials and cluster-randomised trials were excluded, as were small trials with less than 50 patients per group. Any trials with non-random allocation would also have been excluded, but none were identified. We decided that a sample size of 50 trials was large enough to provide representative and reliable results, and was small enough to allow thorough assessment.

Thus 50 trial reports were obtained: 24 in *New England Journal of Medicine*, 15 in *The Lancet*, six in *JAMA*, and five in *BMJ* (see *Lancet* website for references: www.thelancet.com). A standard form detailing the uses of baseline data was piloted on a few trials to achieve consistency of data extraction. Discrepancies were resolved by discussion and agreement, usually when the report was not clear. The data extracted for each report are outlined in the panel.

Results

The 50 randomised trials surveyed had the following characteristics: 39 trials had two randomised treatments, five had three treatments, and six (two with a factorial design) had four treatments. The number of patients

Data extracted from each report

Background
Number of patients
Number of treatment groups
Length of follow-up
Number of centres
Whether primary outcome(s) were predefined
Significance of the overall primary treatment difference

Baseline comparability
Number of baseline factors compared
Any imbalances noted
Whether significance tests were done on baseline data

Randomisation
Practical means of delivering randomised assignments
Statistical method of randomisation
Whether randomisation was balanced by centre
Whether randomisation was balanced by other factors, and if so how many

Covariate adjustment
Whether the primary outcome's results were with or without covariate adjustment
If both, which received more emphasis
Number of covariates adjusted for
Statistical method used
Whether centre-adjusted analysis was done

Subgroup analysis
Whether subgroup analyses were done
Number of subgroup factors
Number of outcomes with subgroup analyses
Number of subgroup analyses (ie, factors × outcomes)
Whether subgroup analyses were pre-planned
Statistical method (descriptive only, subgroup p values, or interaction test)
Whether subgroup differences were found
Whether these differences featured in the summary or conclusions
Overall judgment of whether subgroup analyses were carried out and interpreted appropriately

ranged from 100 to 8803 with a median 494, and 40 trials had multiple centres. Follow-up of patients ranged from 11 days to 15 years (median 1 year). 36 trials had a predefined primary outcome, and 34 trials claimed an overall treatment difference with $p<0·05$, 13 of which had $p<0·001$.

Baseline comparability

Only four trials lacked a table of baseline characteristics by treatment allocation, two of which were previously published. The number of baseline features varied widely with a median 14 features and a maximum 41 features (table 1). The largest table of baseline data occupied nearly a whole journal column.[2]

Half the trials assessed imbalances between treatment groups by significance tests. The investigators of 17 trials reported baseline imbalances. Such declarations of imbalance were based on $p<0·05$ in 12 trials containing 18 significant baseline differences at $p<0·05$, which is 6% of 299 such identifiable significance tests.

Randomisation methods

The statistical method for randomisation was not mentioned in 27 reports (table 2). Most of the rest used randomised permuted blocks within strata.[3] Most multicentre trials did balance randomisation by centre, usually with separate randomised blocks for each centre. Other baseline features were balanced for in many trials, usually for just one or two factors by use of random permuted blocks within strata. The few trials balancing for more factors used minimisation or a similar dynamic method.[4,5]

	Number of trials
Number of baseline variables compared	
0	4
1–4	1
5–9	14
10–19	24
20–29	5
⩾30	2
Significance tests for baseline difference	
Yes	24
No	26
Baseline imbalances noted	
Yes	17
No	33

Table 1: **Baseline variables by treatment group in 50 clinical trials**

The means of delivering randomised assignments was unspecified in 22 trials. Most multicentre trials used contact with a central office (usually by telephone), but use of sealed envelopes was not uncommon.

Covariate adjustment

Most trial reports (38 of 50) emphasised simple outcome comparisons between treatments, unadjusted for baseline covariates (table 3). 14 such trials gave only unadjusted results. The other 24 gave covariate-adjusted results as a back-up to the unadjusted analyses. The remaining 12 reports gave covariate-adjusted analyses primary (or equal) emphasis; six of these 12 gave no unadjusted results. The number of covariates adjusted for varied substantially, with a median three and a maximum 14 covariates. Four reports did not specify the covariates.

The reasons for the choice of covariates were often not clearly explained but we have attempted to formally quantify the varied explanations (table 3). The two main themes were covariates predicting outcome (12 trials, six of which used a stepwise variable-selection procedure),

	Number of trials
Practical means of determining treatment allocation	
Contact with central office	18
Sealed envelopes	8
Blinded packages	2
Unspecified	22
Statistical method used	
Random permuted blocks	15
Minimisation	3
Simple randomisation	1
Other methods	4
Unspecified	27
Balancing for centre	
Yes	18
No	7
Unspecified	14
Single-centre study	11
Balancing for other baseline factors	
Yes	22
No	14
Unspecified	14
Number of baseline factors balanced	
1	8
2	9
3	4
4	0
5	1

Table 2: **Methods of randomisation**

	Number of trials
Were primary outcome analyses done with covariate adjustment?	
No, unadjusted only	14
Yes	36
Which analyses received more emphasis?	
Unadjusted	38
Covariate adjusted	11
Equal emphasis	1
Number of covariates included	
1	7
2	6
3	4
4	2
5–9	11
⩾10	2
Unclear	4
Did covariate adjusted analysis alter the trial conclusions, compared with unadjusted analyses?	
No	29
Yes	1
Only unadjusted given	14
Only adjusted given	6
Reasons for choice of covariates*	
No reason given	15
Covariates were (or expected to be) prognostic	12
Covariates imbalanced between groups	5
Centre or country adjusted for	4
Baseline value of quantitative outcome	3
Other treatment factor in a factorial trial	2
Covariates used in stratified randomisation	1

*More than one reason in some trials.

Table 3: **Covariate adjustment in analysis of patients' response by treatment**

and covariates imbalanced between groups (five trials), reflecting quite different statistical strategies.

The most common statistical methods of covariate adjustment were multiple regression (analysis of covariance) for a quantitative outcome, logistic regression for a binary response, or Cox's proportional hazard models for time-to-event (eg, survival) data. In only one report with unadjusted and covariate-adjusted analyses did the adjustment affect the conclusions.

Subgroup analyses

Most trial reports did include subgroup analyses, that is treatment outcome comparisons for patients subdivided by baseline characteristics (table 4). Many trials confined subgroup attention to just one baseline factor, but five trials examined more than six factors. The number of outcomes subjected to subgroup analysis also varied substantially: many reports studied one outcome for subgroup differences, but six reports explored six or more outcomes.

The total of subgroup analyses is the product of the number of factors and number of outcomes, except for trials with varying baselines factors for different outcomes. The largest number of subgroup analyses was 24 (median four).

Less than half of subgroup-analysis reports used statistical tests of interaction, which directly examine whether the treatment difference in an outcome depends on the patient's subgroup. Most other reports relied on p values for treatment difference in each separate subgroup. A few reports presented subgroup results without statistical tests, simply noting agreement with the overall results. It was commonly difficult to determine whether the subgroup analyses were predefined or post hoc. Most trials lacked power to detect any but very large subgroup effects.

	Number of trials
Were subgroup analyses reported?	
Yes	35
No	15
Number of baseline factors included	
1	17
2	3
3	3
4	5
5	1
6	1
⩾7	5
Number of outcomes for subgroup analysis	
1	17
2	6
3–5	6
⩾6	6
Total number of subgroup analyses	
1	8
2	4
3–5	8
6–8	9
9–11	0
12–24	4
Unclear	2
Statistical method used for subgroup analysis	
Descriptive only	7
Subgroup p values	13
Interaction test	15
Subgroup differences claimed	
Yes	21
No	14
Subgroup claim features in summary or conclusion	
Yes	13
No	8

Table 4: **Subgroup analyses**

Most trials reporting subgroup analysis did go on to claim a subgroup difference, in that the treatment difference depended on the patient's subgroup. Furthermore, most of these claims were stated in the trial's summary or conclusions, or both: 26% of trial reports emphasised the importance of a subgroup finding. Most claims were that the treatment difference was confined to a particular subgroup or was greater in that subgroup. Two claims were in trials in which treatment groups overall did not differ significantly.

Discussion

We have identified some key shortcomings and controversies in the uses of baseline data in clinical trials in current reporting practice.

The CONSORT statement,[1] which is used by many journals for the reporting of controlled trials, does recommend documentation of randomisation methods, but such information is still lacking in many trial reports. Inadequate reporting of randomisation was identified a few years ago[6,7] and as yet there seems little improvement. The achievement of allocation concealment (ie, investigators and patients not knowing the assigned treatment before randomisation takes place) is particularly important,[1,6] but cannot be determined from most trial reports.

When specified, most trials used random permuted blocks, often stratified by one or two baseline factors, and also stratified by centre for multicentre trials.[3] Only a few trials balanced randomisation by more than two factors, but the consequent need for more complex minimisation methods[4,5] is sensibly a common deterrent. Randomisation needs to work reliably with a straightforward delivery of unpredictable random assignments. Having well-balanced treatment groups adds credibility, but the gains in statistical efficiency are negligible.[8] Furthermore, the best predictors of outcome are often not chosen for stratified randomisation. For instance, a trial comparing intervention strategies in angina[9] stratified by two factors unrelated to prognosis, whereas four other strong predictors were subsequently identified. The randomisation process should be as simple and foolproof as possible, only stratifying by centre and factors known to predict outcome.

To show baseline similarity across randomised treatment groups is useful but is often carried to excess. The most useful role of any report's table of baseline data may be an overall description of the characteristics of the patients rather than a comparison of treatment groups. What matters most are the few key predictors of the outcomes of patients, but some authors list many variables in unduly large and unexciting tables. Given journals' restrictions on space for tables and figures, authors may be denying themselves other more interesting displays.

The use of significance tests for detecting baseline differences is questionable.[7] Any differences are either due to chance or to flawed randomisation (a serious bias uncorrectable by statistical analysis). Our 6% rate of significant baseline comparisons agrees with an earlier survey's 4% rate,[7] illustrating nicely that on average 5% of such tests will reach $p<0{\cdot}05$. We agree with Senn[10] and Altman[11] that such significance testing is inappropriate; indeed Senn argues that "this practice is philosophically unsound, of no practical value and potentially misleading". A significant imbalance will not matter if a factor does not predict outcome, whereas a non-significant imbalance can benefit from covariate adjustment if the factor is a strong predictor. But around half of trials still do such significance tests. Some trials gave an extra column of p values in the table of baseline data (eg, Gordin and colleagues[12]), which seems too detailed, whereas others noted just the significant differences. For instance, one report[13] noted significant differences in mean age ($p=0{\cdot}04$) and recent surgery ($p=0{\cdot}02$) while claiming that "baseline characteristics of the patients were similar in the two groups". For the primary outcome the investigators reported that adjusting for these two factors did not alter the results. No information was given on whether these or other factors were associated with outcome, a more interesting insight than the focus on unlucky significant differences.

Reports vary enormously in their use of covariate-adjusted analyses, and this merits clearer guidelines.[14] The good news is that only one trial surveyed found a difference between unadjusted and covariate-adjusted analysis sufficient to affect the conclusions; most estimates (eg, mean difference, relative risk), confidence limits, and p values were very similar. The likely explanation is that most covariates are not strongly related to outcome and are well-balanced between treatments. The one exception[15] was a trial of cryptococcal meningitis in which the treatment difference in risk of positive cerebrospinal-fluid culture after 2 weeks had unadjusted odds ratio of 1·47 ($p=0{\cdot}06$) and odds ratio of 1·92 ($p=0{\cdot}01$) after adjustment for three predictive covariates. However, covariate adjustment excluded 31% of patients because of missing covariate information, which casts doubt on the reliability of such an analysis.

When could covariate adjustment be important? Senn[16] uses Normal theory to show how the reliability of an unadjusted significance test is affected by both the correlation coefficient r between the covariate and outcome, and the covariate's standardised treatment imbalance z. He shows that non-significant covariate imbalance can matter if the covariate is strongly related to outcome. For instance, a trial of primary biliary cirrhosis[17] had a non-significant imbalance in a strongly prognostic variable, serum bilirubin. Unadjusted and adjusted analyses gave $p=0{\cdot}2$ and $p=0{\cdot}02$, respectively, for the treatment differences in survival. Conversely, if the correlation is weak (eg, $r\leqslant0{\cdot}1$), then even a significant covariate imbalance has little impact on the validity of the unadjusted analysis.

Although many covariates have significant associations with outcome, few achieve sizeable correlations.[18] However, the baseline value of a quantitative outcome often has correlation $r>0{\cdot}5$.[19] Covariate adjustment for such a baseline achieves improved precision of estimation and more valid significance testing. Analysis of covariance is better than ignoring the baseline or overcorrecting by taking differences from baseline.[20]

For logistic regression[21] and proportional hazards models[22] covariate-adjusted estimates are not more precise, but the odds ratio or hazard ratio becomes further away from the null. Adjustment for strong predictors of outcome achieves more relevant treatment-effect estimates and significance tests.

Researchers commonly cannot predeclare the strong predictors, so the choice of a covariate-adjusted analysis is determined by a variable selection procedure. Such data-driven covariate adjustment can arouse suspicion (eg, did investigators favour their conclusions by this particular covariate choice?), which adds to the argument that emphasis should be on simple unadjusted analyses.

Covariate-adjusted results are also harder for readers to understand. For instance, one study[23] presented rate-group differences on a logarithmic scale adjusted for eight covariates. Another[24] presented adjusted results of a mixed linear model for repeated measures, without any unadjusted analyses. Readers need more help to understand such complexities.

Should one adjust for centre (or country) in multicentre (or international) trials? This adjustment probably makes no difference but helps to confirm a primary unadjusted analysis. Investigating treatment-by-centre interactions usually lacks statistical power, but as a quality check can reveal concerns over centres with exceptionally large or small treatment differences.

Of all the various multiplicity problems in clinical trials[25] subgroup analysis remains the most overused and overinterpreted. The problems have been elucidated[26,27] but reports show a continued desire to undertake many subgroup analyses. This reflects the intellectually important issue that real treatment differences may depend on certain baseline characteristics.

Thus, such data exploration is not bad on its own provided investigators (and readers) do not overemphasise subgroup findings. Also, the underuse of statistical tests of interaction means the play of chance gets inadequate recognition. Reliance on subgroup p values is misleading. If the overall result is significant, almost inevitably some subgroups will and some will not show significant differences depending on chance and the smallness of subgroups. Conversely, if an overall difference is not significant, some subgroups may have a bigger observed treatment difference by chance, which may even reach significance. For instance, an intervention trial after myocardial infarction[28] found no overall mortality difference, but gave much space to separate analyses and conclusions for men and women, because the cardiac mortality difference for women seemed greater (subgroup p=0·06). The summary gave results separately for men and women, inferring "the possible harmful impact of the intervention on women". The statistical interaction test would have helped: it assesses whether 22 versus 12 deaths for women (odds ratio 2·0) is significantly different from the 11 versus 11 deaths for men (odds ratio 1·0). Interaction test p=0·21 indicates insufficient evidence that the intervention's effect (if any) depended on sex.

As in this case, interaction tests commonly lack statistical power (ie, the trial is not large enough to detect subgroup findings). Hence one is inevitably left in doubt as to whether a suggestive subgroup analysis (eg, with interaction p=0·05 to p=0·15) is simply due to chance or merits further investigation. Even after reporting a non-significant interaction test, investigators may still overinterpret subgroup findings. For instance, one trial had an overall highly significant 43% relative reduction in risk of heart failure.[29] For patients with previous myocardial infarction this increased to 76%. This subgroups' numbers of heart failure events were small (five *vs* 17) and the interaction test was not significant (p=0·24), but the report still emphasised the finding in the summary, which concluded that "amongst patients with previous myocardial infarction, an 80% risk reduction was observed".

Both these reports illustrate how time-to-event plots comparing treatments by subgroups mislead one into exaggerating the evidence of a subgroup effect. The problem is such plots usually do not display the data's statistical uncertainty, lacking standard errors or confidence limits.

When the interaction test is significant, how much emphasis should the subgroup findings receive? In an angina trial[8] the highly significant overall treatment differences in angina grade and exercise time 6 months after randomisation were absent in patients with less severe angina or good exercise time at baseline. Although the four interaction tests were highly significant, five other baseline factors had also been assessed for subgroup effects. Hence the summary gave only overall results, adding the phrase "especially in patients with more severe angina" as a marker of exploratory subgroup findings.

The BARI trial,[30] which compared coronary artery bypass grafting and percutaneous transluminal coronary angioplasty in angina, illustrates the dilemma of what to do when a highly significant but unexpected interaction is found. There was no overall survival difference, but in diabetic patients mortality was nearly double in the angioplasty group compared with the bypass-graft group (interaction test p=0·003). The investigators were clearly convinced by this finding, which led to a major public-health recommendation by the US National Institutes of Health. However, this was one of several exploratory subgroup analyses so that the risk of an exaggerated false positive is not negligible. We feel that such a surprise subgroup finding should have been a basis for further research (eg, from other similar trials), rather than an immediate influence on national policy.

The good judgment of investigators, referees, and editors determines what emphasis subgroup analyses should get. Questions about which types of patients benefit most from a new treatment are clinically important, but most studies lack the statistical power to identify such subgroup effects. Even with prespecified subgroup analyses, post-hoc emphasis on the most fascinating subgroup finding inevitably leads to exaggerated claims. We suspect that some investigators selectively report only the more interesting subgroup analyses, thereby leaving the reader (and us) unaware of how many less-exciting subgroup analyses were looked at and not mentioned.

On the whole, our survey indicates that subgroup analyses occupy much space in clinical trial reports, and influence conclusions more often than is justified.

Recommendations

Randomisation methods

Randomisation procedures, both the practical means of treatment allocation and the statistical methods used, need clearer explanation. In particular, reports should state which baseline factors the randomisation balanced for, and by what method. In design, balancing should be confined to centre and factors known to be strong predictors of outcome.

Baseline comparisons

Although reports should show in appropriate detail the types of patient included, the baseline comparisons across treatments need not be so extensive. The table of baseline comparability should mainly focus on baseline factors thought to be associated with primary outcomes.

Significance tests for baseline differences are inappropriate. A chance significant baseline imbalance is unimportant if the factor is unrelated to outcome, unless it signals errors in randomisation. Conversely, if a baseline factor strongly influences outcome, a non-significant treatment imbalance may be important.

Covariate adjustment

In general, simple unadjusted analyses that compare treatment groups should be shown. Indeed they should be emphasised, unless the baseline factors for covariate adjustment are predeclared on the basis of their known strong relation to outcome. One notable exception is the baseline value of a quantitative outcome, in which analysis of covariance adjustment is the recommended primary analysis since a strong correlation is expected.

Many trials lack such prior knowledge, requiring any strong predictors of outcome to be identified from the trial data by use of an appropriate variable selection technique. Covariate adjustment should then be a secondary analysis. Adjustment for baseline factors with treatment imbalances is unimportant, unless such factors relate to outcome. Nevertheless, such secondary analyses help achieve peace of mind.

Covariate-adjusted analyses are more complicated, but investigators should ensure that readers can understand them. Increased technicality should not be an excuse for lack of clarity.

Subgroup analysis

Investigators should be cautious when undertaking subgroup analyses. Subgroup findings should be exploratory, and only exceptionally should they affect the trial's conclusions. Editors and referees need to correct any inappropriate, overenthusiastic uses of subgroup analyses.

The credibility of subgroup analyses is improved if confined to the primary outcome and to a few predefined subgroups, on the basis of biologically plausible hypotheses. This might include factors used to stratify randomisation. Investigators should recognise whether their trial is not large enough to detect realistic subgroup effects, and be particularly wary of claiming a treatment difference in a subgroup when the overall treatment comparison is not significant. Such subgroup rescues of otherwise negative trials are often unwarranted, unless the evidence is statistically convincing and clinically sensible.

Statistical tests of interaction (that assess whether a treatment effect differs between subgroups) should be used rather than inspection of subgroup p values, which often encourages inappropriate subgroup claims. Only if the statistical interaction test supports a subgroup effect should the conclusions be influenced. Even then, the emphasis should depend on biological plausibility, the number of subgroup analyses, their prespecification, and the statistical strength of evidence, recognising that most subgroup claims are prone to exaggerate the truth.

In multicentre trials, centre-adjusted analysis and treatment-by-centre interactions may be useful secondary analyses, but should not replace the overall results.

Conclusion

Clinical trial reports need a clearly defined policy on uses of baseline data, especially with respect to covariate adjustment and subgroup analysis. There are substantial risks of exaggerated claims of treatment effects arising from post-hoc emphases across multiple analyses. Subgroup analyses are particularly prone to overinterpretation, and one is tempted to suggest "don't do it" (or at least "don't believe it") for many trials, but this suggestion is probably contrary to human nature.

Contributors

Susan Assmann, Laura Enos, and Linda Kasten collaborated in study design and in writing the paper, extracted data from the reports, and tabulated results. Stuart Pocock wrote the paper, devised the study, oversaw its planning and execution, and resolved discrepancies.

References

1 Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomised controlled trials: the CONSORT statement. *JAMA* 1996; **276:** 637–39.

2 Sundberg K, Bang J, Smidt-Jensen S, et al. Randomised study of risk of fetal loss related to early amniocentesis versus chorionic villus sampling. *Lancet* 1997; **350:** 697–703.

3 Friedman LM, Furberg CD, DeMets DL, eds. The randomisation process. In: Fundamentals of clinical trials, 3rd edn. New York: Springer-Verlag, 1998: 61–78.

4 Birkett JJ. Adaptive allocation in randomised controlled trials. *Control Clin Trials* 1985; **6:** 146–55.

5 Freedman LS, White SJ. On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics* 1976; **32:** 691–94.

6 Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomisation from reports of controlled trials published in obstetrics and gynecology journals. *JAMA* 1994; **272:** 125–28.

7 Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990; **335:** 149–53.

8 Pocock SJ, Lagakos SW. Practical experience of randomisation in cancer trials: an international survey. *Br J Cancer* 1982; **46:** 368–75.

9 RITA-2 trial participants. Coronary angioplasty versus medical therapy for angina: the second randomised intervention treatment of angina (RITA-2) trial. *Lancet* 1997; **350:** 461–68.

10 Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994; **13:** 1715–26.

11 Altman DG. Comparability of randomised groups. *Statistician* 1985; **34:** 125–36.

12 Gordin FM, Matts JP, Miller C, et al. A controlled trial of isoniazid in persons with anergy and human immunodeficiency virus infection who are at high risk for tuberculosis. *N Engl J Med* 1997; **337:** 315–20.

13 Columbus Investigators. Low-molecular-weight heparin in the treatment of patients with venous thromboembolism. *N Engl J Med* 1997; **337:** 657–62.

14 Altman DG. Adjustment for covariate imbalance. In: Armitage P, Colton T, eds. Encyclopaedia of biostatistics. Chichester: John Wiley, 1998: 1000–05.

15 Van der Horst CM, Saag MS, Cloud GA, et al. Treatment of cryptococcal meningitis associated with the acquired immunodeficiency syndrome. *N Engl J Med* 1997; **337:** 15–21.

16 Senn SJ. Covariate imbalance and random allocation in clinical trials. *Stat Med* 1989; **8:** 467–75.

17 Christensen E, Neuberger J, Crowe J, et al. Beneficial effects of azathioprine and prediction of prognosis in primary biliary cirrhosis: final results of an international trial. *Gastroenterology* 1985; **89:** 1084–91.

18 Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Control Clin Trials* 1989; **10:** 161S–75S.

19 Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med* 1992; **2:** 1685–704.

20 Snedecor GW, Cochran WG, eds. Analysis of covariance. In: Statistical methods. Ames: Iowa State University Press, 1989: 419–46.

21 Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991; **59:** 227–40.

22 Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med* 1995; **14:** 735–46.

23 Olds DL, Eckenrode J, Henderson CR, et al. Long-term effects of home visitation on maternal life course and child abuse and neglect. *JAMA* 1997; **278:** 637–43.

24 Rolfs RT, Riduan Joesoef M, Hendershot EF, et al. A randomised trial of enhanced therapy for early syphilis in patients with and without human immunodeficiency virus infection. *N Engl J Med* 1997; **337:** 307–14.

25 Pocock SJ, ed. Multiplicity of data, section 14.3. In: Clinical trials: a practical approach. Chichester: Wiley, 1983.

26 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. *JAMA* 1991; **266:** 93–98.

27 Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *N Engl J Med* 1987; **317:** 426–32.

28 Frasure-Smith N, Lespérance F, Prince RH, et al. Randomised trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction. *Lancet* 1997; **350:** 473–79.

29 Kostis JB, Davis BR, Cutler J, et al. Prevention of heart failure by antihypertensive drug treatment in older persons with isolated systolic hypertension. *JAMA* 1997; **278:** 212–16.

30 Bypass Angioplasty Revascularisation Investigation (BARI) Investigators. Comparison of coronary bypass surgery with angioplasty in patients with multivessel disease. *N Engl J Med* 1996; **335:** 217–25.