Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems

Stuart J. Pocock[1,*,†], Susan E. Assmann[2], Laura E. Enos[2] and Linda E. Kasten[2]

[1]*Medical Statistics Unit, London School of Hygiene & Tropical Medicine, London, WC1E 7HT, U.K.*
[2]*New England Research Institutes, Watertown, MA 02472, U.S.A.*

SUMMARY

Clinical trial investigators often record a great deal of baseline data on each patient at randomization. When reporting the trial's findings such baseline data can be used for (i) *subgroup analyses* which explore whether there is evidence that the treatment difference depends on certain patient characteristics, (ii) *covariate-adjusted analyses* which aim to refine the analysis of the overall treatment difference by taking account of the fact that some baseline characteristics are related to outcome and may be unbalanced between treatment groups, and (iii) *baseline comparisons* which compare the baseline characteristics of patients in each treatment group for any possible (unlucky) differences. This paper examines how these issues are currently tackled in the medical journals, based on a recent survey of 50 trial reports in four major journals. The statistical ramifications are explored, major problems are highlighted and recommendations for future practice are proposed. Key issues include: the overuse and overinterpretation of subgroup analyses; the underuse of appropriate statistical tests for interaction; inconsistencies in the use of covariate-adjustment; the lack of clear guidelines on covariate selection; the overuse of baseline comparisons in some studies; the misuses of significance tests for baseline comparability, and the need for trials to have a predefined statistical analysis plan for all these uses of baseline data. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: clinical trials; subgroup analysis; covariate adjustment; baseline comparisons; medical journals

1. INTRODUCTION

Clinical trials usually entail the recording of substantial amounts of baseline data on each patient at randomization. These data document the patient's current medical condition (for example, signs, symptoms, quantitative measures, ancillary medications), medical history (for example, previous disease events, time since diagnosis) and demographics (for example, age, sex and other personal characteristics).

---

* Correspondence to: Stuart Pocock, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.
† E-mail: stuart.pocock@lshtm.ac.uk

When a trial's results are reported such baseline data have three main uses: (i) *subgroup analyses*, whose purpose is to examine if any treatment differences in patients outcome (or lack thereof) appear consistent across all types of patients or depend on one or more baseline variables; (ii) *covariate-adjusted* analyses, whose purpose is to take account of any baseline variables that are related to patient outcome (especially if they are not balanced across treatment groups) in order to achieve the most reliable statistical estimates and tests for the overall treatment differences in outcome; (iii) *baseline comparisons*, whose purpose is to document the types of patients in the study and to demonstrate the extent to which the treatment groups were similar prior to commencement of randomized treatment.

This paper aims to briefly describe the key statistical issues that are pertinent to high quality standards of reporting for these three topics. In each case such desirable standards are compared with current reporting practice by means of a survey of recent trial reports in major medical journals. The survey is reported more fully elsewhere [1]. It comprised all reports of parallel group clinical trials in which over 50 patients per treatment group were individually randomized and which were published in the *British Medical Journal* (*BMJ*), the *Journal of the American Medical Association* (*JAMA*), the *Lancet* and the *New England Journal of Medicine* (*NEJM*) during July to September 1997. Fifty trial reports were thus obtained: 24 in the *NEJM*, 15 in the *Lancet*, 6 in the *JAMA* and 5 in the *BMJ*. A pre-piloted standard form detailing each report's uses of baseline data was filled out by three of the authors (SFA, LEE and LEK) with any discrepancies resolved by consensus across all four authors.

Section 2 addresses subgroup analyses, integrating desirable statistical practice with the reality of current practice in the journals in order to formulate some key recommendations. Similarly, Sections 3 and 4 investigate covariate-adjustment and baseline comparability respectively. Section 5 sums up overall.

2. SUBGROUP ANALYSIS

Patients recruited into a clinical trial are not a homogeneous sample. Their response to treatment and the differing impact on them of different treatments may well vary in ways that affect the choice of which treatment is best for which patient. Thus, if in truth there are specific subgroups of patients for which a new treatment is more (or less) effective (or harmful) than is indicated by the overall comparison with standard treatment in the trial as a whole, we have a scientific and ethical obligation to try and identify such subgroups. As a consequence most trial reports ($35/50 = 70$ per cent in our survey sample, see Table I) do contain some results of subgroup analyses.

Several difficulties arise though when undertaking subgroup analyses [2, 3]:

(i) Most trials only have sufficient statistical power (if that) to detect the overall main effect difference in response between treatment groups, so that if subgroup effects do exist, they may well go undetected because the trial was not large enough. Indeed, most trials we surveyed could only have detected very large subgroup effects.

(ii) Given the plethora of baseline variables and the tendency not to have a clear predefinition of which subgroup(s) may be more (or less) differentially responsive to a new treatment, there are many possible subgroup analyses that could be performed. Hence one needs to guard against data dredging and the potential for *post hoc* emphasis on the 'most interesting' across many subgroup analyses.

Table I. Subgroup analyses in 50 clinical trial reports.

		Number of trials
Were subgroup analyses reported?	Yes	35
	No	15
Number of baseline factors included	One	17
	Two	3
	Three	3
	Four	5
	Five to six	2
	Seven or more	5
Number of outcomes for subgroup analysis	One	17
	Two	6
	Three to five	6
	Six or more	6
Total number of subgroup analyses	One	8
	Two	4
	Three to five	8
	Six to eight	9
	12 to 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
If yes:		
Subgroup claim featured in abstract and/or conclusions	Yes	13
	No	8

For instance, in our survey only eight trials (16 per cent) reported just one subgroup analysis. It was common to examine more than one baseline factor, and also to study each subgroup categorization for several different outcome variables. The total number of reported subgroup analyses (that is, number of baseline variables times number of outcomes) varied enormously with a maximum of 24 subgroup analyses and a median of four subgroup analyses amongst the 35 trials with any at all. Of course, this does not include any further unreported subgroup analyses which authors may also have carried out. In all but a few trials it was not possible to determine whether the subgroup analyses were according to a predefined statistical analysis plan or arose from *post hoc* data exploration.

(iii) The most appropriate statistical methods for making inferences from subgroup analyses are often not used in trial reports. In our view, statistical tests for interaction, which directly examine the strength of evidence for the treatment difference varying between subgroups, are the most useful approach for evaluating subgroup analyses. Sometimes

the fact that interaction tests usually lack statistical power is put forward to argue against their use. However, we feel this is the very reason they are of great value: interaction tests recognize the limited extent of data available for subgroup analysis, and are the most effective statistical tool in inhibiting false or premature claims of subgroup findings. In the survey, only 15 (43 per cent) of the 35 reports with subgroup analyses used interaction tests; 13 reports (37 per cent) instead presented *p*-values for treatment difference in each separate subgroup, but subgroup *p*-values can be misleading. If the overall treatment difference is statistically significant, then it is very likely that some subgroups will and some will not show a significant treatment difference depending on chance and the smallness of subgroups. A further seven reports simply presented the subgroup findings without any statistical tests, observing that all subgroup analyses were consistent with the overall result. Had any subgroup inconsistencies been apparent one assumes some more formal inference, preferably an interaction test, would have been carried out.

(iv) The extent to which subgroup analyses should affect the interpretation and conclusions in a trial report is a contentious matter. While responsible triallists need to conclude whether a treatment effect (or lack of effect) is not generalizable to certain type(s) of patient, they also need to guard against making exaggerated subgroup claims that are not sufficiently robust to affect treatment policy. In our survey, 21 trial reports (42 per cent) claimed to find subgroup differences that appeared incompatible with the overall treatment comparison, and 13 of these went on to feature such claims in the summary and/or conclusions. These claims were mostly that the treatment difference existed only in a particular subgroup (what might be called an 'all or nothing' interaction), or was more marked in a particular subgroup (commonly called a 'quantitative' interaction). So called 'qualitative' interactions in which the treatment effect is in opposite directions in different subgroups are thought to be rare and highly implausible, and indeed did not occur in this survey.

In general, once the statistical strength of evidence for interaction is documented correctly, one relies on the wise judgement of the triallists (and also journal referees and editors) in deciding what emphasis any subgroup finding should receive. Both our survey and other experiences lead us to the view that at present subgroup analyses are overinterpreted by authors (and probably readers as well) and that much greater caution needs to be exercised when drawing conclusions on subgroups. Biological plausibility, the number of subgroup analyses performed, their prespecification and the trial's size all need to be considered alongside the statistical strength of evidence when weighing up the all too likely case that any particular subgroup finding, no matter how intriguing, is prone to be an exaggeration of the truth. In this regard Bayesian strategies for subgroup analysis [4–6], including tests of qualitative interactions, are an interesting development.

Many of the above issues in subgroup analysis are encapsulated in two examples from our survey, whose survival plots for the subgroups they focused on are displayed in Figure 1. Such time-to event survival curves comparing treatment by subgroups are particularly prone to accentuate suggestions of a subgroup effect, because they do not present the data's statistical uncertainty by including standard errors or confidence limits.

The first example concerns the suggestion that psycho-social nursing intervention after myocardial infarction appears to adversely affect cardiac mortality in women but not in men [7]. The article's summary accentuated this point by only giving results separately for each
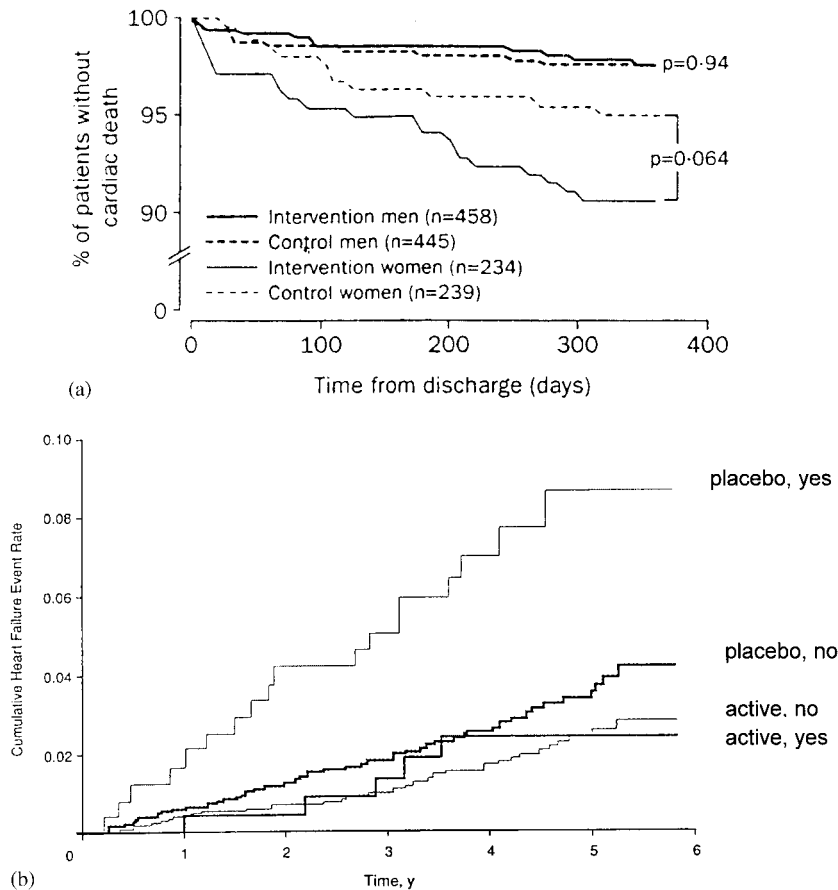
Figure 1. Examples of subgroup analyses with survival plots. (*a*) Cardiac death by gender in a study of psycho-social nursing intervention after myocardial infarction. (*b*) Heart failure by previous myocardial infarction (yes or no) in a trial of antihypertensive treatment (active versus placebo).

gender, and the concluding sentence states 'the possible harmful impact of the intervention on women'. The subgroup *p*-value for the treatment difference in women, $p = 0.064$ in Figure 1(*a*), helped to maintain this interest, whereas the following more appropriate interaction test would have encouraged a more cautious perspective: there were 22 versus 12 deaths in women (odds ratio 2.0) but 11 versus 11 deaths in men (odds ratio 1.0) and the interaction test comparing these odds ratios has $p = 0.21$, demonstrating the lack of evidence that the intervention's effect (if any) on cardiac mortality depended on sex.

The second example [8], in Figure 1(*b*), suggests that antihypertensive treatment reduces the risk of heart failure more markedly in patients with a history of myocardial infarction. The observed relative risk reductions for patients with and without such a history are 76 per cent and 33 per cent, respectively, and the wide 95 per cent confidence interval for the former (18 to 92 per cent reduction) reflects the fact that it is based on only 5 versus 17 heart failure occurrences. Accordingly, the authors' inclusion of the statistical interaction test ($p = 0.24$) in the Results section, would appear to justify a de-emphasis of this subgroup finding, out of

the many that could have been undertaken. However, the article's summary drew particular attention to this subgroup, the final sentence of conclusions stating 'amongst patients with prior MI, an 80 per cent risk reduction was observed', whereas the overall risk reduction of 43 per cent (with 95 per cent CI 19 to 66 per cent) would seem a more appropriate summary statistic.

3. COVARIATE ADJUSTMENT

Experience shows that for most clinical trials, analyses which adjust for baseline covariates are in close agreement with the simpler unadjusted treatment comparisons. This is because (a) the randomization usually results in well balanced treatment groups, and (b) most covariates are not strongly related to the outcome.

Nevertheless, the statistical properties of covariate-adjustment are quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy. The primary aim is to achieve an unbiassed and statistically efficient treatment comparison which takes account of baseline factors that predict prognosis, especially those factors that have some imbalance between treatment groups. In addition, there is some credibility attached to demonstrating that covariate adjustment does not alter the conclusion derived from unadjusted analyses. A further benefit can be the creation of a predictive model which combines the influences of treatment and prognostic covariates in estimating the expected outcome (and absolute treatment benefit) for individual patients.

Problems arise in the selection of the covariates. Consideration may be given to baseline factors that (i) predict outcome, possibly using a stepwise variable selection algorithm, (ii) are imbalanced between groups (but using what criterion?) and/or (iii) were used to stratify the randomization, and some would also argue that only covariates that are prespecified in the protocol or statistical analysis plan may be permitted. The scope for judgements in an ill-defined strategy, and biased (for example, most favourable) choices out of a multiplicity of possible analyses, means that covariate adjusted analyses may rightly be viewed with some suspicion, often leaving primary emphasis on the unadjusted analysis.

Therefore, let us review some of the statistical properties of covariate adjustment [9–11]. There can be several statistical aims:

 (i) to achieve the most appropriate *p*-value for the treatment difference;
 (ii) to achieve an unbiassed estimate and confidence interval for the magnitude of treatment difference in outcome;
(iii) to improve the precision of the estimated treatment difference, thus increasing the statistical power of the trial.

Let us first consider these issues from the idealized situation of two treatments, a Normal response with known variance and a single covariate also with known variance, as previously explored by Senn [11].

Let Z_x be the standardized imbalance between treatment groups for this covariate x. Under an unstratified randomization scheme Z_x would follow a standardized Normal distribution, whereas any stratification on x or other variables correlated with x would constrain Z_x to have variance smaller than 1. However, what matters here is the observed value of Z_x in this particular trial. Let ρ be the correlation between the covariate and outcome within each treatment.
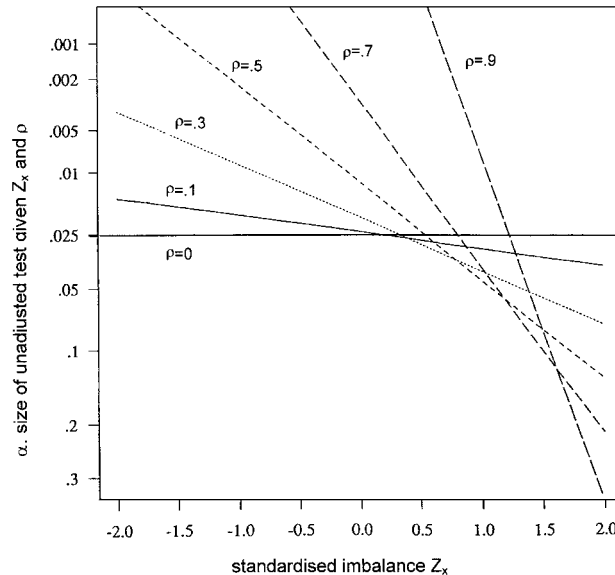
Figure 2. The effect of standardized covariate imbalance Z_x and the covariate's correlation with outcome ρ on the conditional size of unadjusted one-sided test (true $\alpha = 0.025$).

Achieving the correct p-value in any trial is often considered important, especially in pharmaceutical company sponsored trials with a regulatory implication. Conditional on the observed value of Z_x, an analysis of covariance adjusting for x will generate an appropriate one-sided p-value, that is, prob (ANCOVA p-value $< \alpha \,|\, Z_x, H_0) = \alpha$, for any choice of α such as 0.025 for two-sided 5 per cent significance. Now, this is not so for an unadjusted test of the treatment difference in mean outcomes since it ignores the treatment imbalance in a baseline covariate correlated with the outcome. From Senn, the size of the unadjusted one-sided test, conditional on the observed Z_x, is as follows:

$$\text{prob (unadjusted } p\text{-value} < \alpha \,|\, Z_x, H_0)$$
$$= \text{prob} \left(\text{standardized normal deviate} > \frac{Z_\alpha - Z_x \rho}{\sqrt{\{1 - \rho^2\}}} \right)$$

where Z_α is the standardized normal deviate associated with a one-tail probability α.

We dwell on a one-sided test, because the direction of treatment imbalance affects the two tails of a two-sided test in contrasting ways. Figure 2 illustrates this impact of Z_x and ρ on desired $\alpha = 0.25$, the most commonly chosen size of such an unadjusted one-sided test. First, for $\rho = 0$ the size is unaffected no matter how great the imbalance in x. For a strongly correlated covariate (for example, with $\rho = 0.7$) the impact of Z_x is marked. For instance, with $Z_x = \pm 1.5$, the unadjusted α becomes 0.102 in one direction and < 0.0001 in the other direction. Note that even for perfect balance, that is, $Z_x = 0$, the unadjusted test has become markedly conservative (unadjusted $\alpha = 0.003$) due to the fact that perfect balance in a strong predictor is constraining the outcome variability between treatment groups. In practice, a correlation as high as 0.7 is quite plausible for the same variable measured at baseline and

　　　　　　　　Statist. Med. 2002; **21**:2917–2930

after treatment [12] and the above illustrates one reason why ANCOVA adjusting for this baseline is crucially important.

In our experience, other baseline covariates, whether quantitative or binary, generally have much weaker correlations with outcomes, in which case adjustment for such covariates has less impact on the size of the unadjusted test. For instance, for $\rho = 0.1$, even a statistically significant covariate imbalance such as $Z_x = \pm 2.0$, leads to conditional $\alpha = 0.015$ and 0.038 in the two tails, neither differing much from the desired $\alpha = 0.025$.

The practical conclusion here is that if the correlation is weak, for example, $\rho < 0.3$, even a statistically significant covariate imbalance is unimportant (except as an indicator that the randomization may have been performed incorrectly). On the other hand, if a covariate is strongly related to outcome, for example, $\rho > 0.5$, then it is important to adjust for it regardless of the extent (or lack) of covariate imbalance.

Incidentally, these results are independent of sample size, and add weight to the argument that appropriate covariate adjustment is still important in large trials. However, some would argue that such obsession with significance testing is not what really matters and considerations of *precision* and *bias* in estimates of mean treatment difference are more important.

For the same simple idealized Normal known variance model used above, the ratio of standard errors for the covariate-adjusted and unadjusted estimates of treatment difference = $\sqrt{(1 - \rho^2)}$. Thus, if $\rho = 0.7$ the width of the confidence interval is reduced by a substantial 29 per cent, whereas for $\rho = 0.1$ the reduction is only 0.5 per cent. Consequently, with a single predefined covariate, for ANCOVA to achieve the same statistical power as an unadjusted analysis, the required sample size is reduced proportionately by $1 - \rho^2$. Thus with a very strong predictor with $\rho = 0.7$, such as may occur with a baseline measure of the outcome, the required number of patients is roughly halved. The saving is less than 10 per cent, for $\rho = 0.3$, a value which is not exceeded for most baseline variables. Incidentally, when the same measure is obtained at baseline and after treatment, ANCOVA also wins over an analysis based on changes [13], the ratio of standard errors (ANCOVA versus changes) then being $\sqrt{\{(1 + \rho)/2\}}$.

Conditional on the observed Z_x, the unadjusted estimate of treatment difference is biased by ρZ_x standard errors. This gets smaller with increased sample size, but as noted earlier the impact on type I error does not change with sample size.

This idealized model for a single covariate under Normal theory usefully quantifies the key desirable consequences of covariate adjustment. However, for binary or survival outcomes, using logistic or proportional hazard models, respectively, the statistical properties of covariate adjustment are rather different [14–16]. The covariate-adjusted estimates are not made more precise (in fact the standard error tends to increase slightly) but more importantly the covariate-adjusted estimates, for example, of odds ratio or hazard ratio, are further from the null. That is, the unadjusted analysis tends to dilute the impact of treatment by failing to compare like with like. For instance, at its simplest consider a binary outcome and a binary covariate strongly related to outcome. The Mantel–Haenszel estimator of the odds ratio comparing treatments is then a more focused, more statistically powerful estimate of treatment effect than the crude odds ratio from a single 2 by 2 table ignoring the covariate. Amidst these complexities, the good news is that the other two benefits of covariate-adjustment, achieving the correct size of test and increasing statistical power, apply equally well to binary and time-to-event outcomes.

With multiple covariates, both quantitative and binary, the same issues apply, but with the added complication of which variables to choose. Simulation studies [17, 18] have shown how *post hoc* selection of covariates for adjustment out of a larger set of potential covariates will tend to lead to biased estimates of the treatment effect, especially with small studies. However, this particular risk has been somewhat exaggerated and the extent to which this is a serious problem in reasonably large trials has not been adequately explored to date. A more contentious issue arises if the variable selection procedure is not totally objective, which allows investigators to focus on the covariate model that best accentuates the estimate and/or statistical significance of the treatment difference.

Some have argued that prespecification of all covariates for adjustment in a single model is the best solution, leading to just one predefined covariate-adjusted analysis [10]. While desirable in principle, this is often unachievable in practice. In many trials, one has inadequate prior knowledge as to which baseline factors are related to prognosis. Consequently any prespecified 'mandatory' list of covariates will often include some irrelevant factors and exclude some powerful predictors. Therefore, in real life we see a continuing need for defining objective variable selection algorithms which lead to the most appropriate covariate-adjusted model. While this means the chosen covariates themselves could not be prespecified, a precise predefined statistical strategy for variable selection should overcome somewhat any suspicions that *post hoc* selection of covariates might be based on subjective criteria.

Another issue is whether one should automatically adjust for variables used to stratify the randomization. Some say definitely 'yes' [10], but this is not necessarily sensible, especially if the covariate(s) in question are not actually related to outcome.

Whether and how to adjust for centre in a multi-centre trial is another complex issue, which requires more space than is possible here. Briefly, our experience suggests it rarely makes any difference, there is a lack of clear guidelines on what to do about small centres or lots of centres or aggregation of centres, but nevertheless demonstration that centre-adjusted analyses agree with the unadjusted analyses can help the credibility of the latter.

Now, let us return to the survey of trial reports, and see how covariate-adjustment is being used in practice. Table II presents the main findings; 36 reports (72 per cent) did include covariate-adjusted analyses, but mostly as a secondary back-up to the unadjusted analyses. Only in 12 reports did the covariate-adjusted analyses get primary (or equal) emphasis. This focus on unadjusted analyses may arise for several reasons: (i) authors and readers prefer the simplicity and clarity of unadjusted findings; (ii) suspicions regarding the potential manipulations of data-driven covariate-adjustment make them less credible; (iii) covariate-adjustment rarely makes much difference so why make the conclusions more complicated; (iv) in some trials, there is insufficient clinical agreement or interest in considering which covariates should (or could) be adjusted for.

Trials varied substantially as regards the number of covariates adjusted for; some chose just one covariate, but it was quite common to include five or more covariates and in a few trials the adjustment was so unclear that the number was unknown.

The reasons underpinning the choice of covariates were often not given, but when known the two main reasons were:

 (i) covariates that predict outcome, 12 trials, six of which adopted a stepwise variable-selection procedure;
(ii) covariates imbalanced between treatment groups, five trials.

Table II. Covariate adjustment in analysis of patient response by treatment.

		Number of trials
Were primary outcome analyses done using 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	One	7
	Two	6
	Three	4
	Four	2
	Five to nine	11
	Ten or more	2
	Unclear	4
Did covariate adjusted analysis alter the trial conclusions, compared to 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 randomization	1

*More than one reason in some trials.

In only one report [19] did the conclusions alter as a result of covariate-adjustment. This seems to have arisen because nearly one-third of patients were excluded because of missing covariate data, surely an inappropriate analysis.

In three trials the baseline value of a quantitative outcome was adjusted for. As mentioned above the strong correlation present here mandates that ANCOVA is indeed the appropriate analysis [12, 13]. However, informal inspection of the same four journals this year (2000) indicates that some reports are still failing to use ANCOVA, either ignoring the baseline or analysing changes instead.

A further problem evident from our survey is that some reports present the covariate-adjusted results in a manner that will be hard to understand for many readers. For instance, one report [20] gave group differences in event rates on a logarithmic scale, adjusted for eight covariates, whereas transformation back to geometric means (or avoidance of the log transform altogether) would have been more helpful. Another report [21] undertook a mixed linear model for repeated measures with both baseline and time-dependent covariates, which surely only a small minority of readers could have followed. Thus, more effort needs to be put into making covariate adjustments better described and more comprehensible.

Overall, what matters is to adjust for the appropriate covariates (that is, the strong predictors of outcome) and to make one's statistical policy for covariate adjustment completely objective. While one can of course recommend prespecification of the specific covariates to adjust for, this will be unrealistic in many instances. While variable selection procedures could be manipulated and may lead to biased estimates in smaller trials, we still feel they have a useful role in formulating covariate-adjustment in larger trials, and wish to encourage more methodological research on this topic.

For peace of mind and credibility some reports will doubtless continue to adjust for irrelevant covariates (for example, those used in stratified randomization or unbalanced between groups, but which are not related to outcome), but at least no harm is done by such statistical excesses.

4. BASELINE COMPARABILITY

In most trial reports Table I is devoted to comparing the distributions of several baseline variables by treatment group. The aims of this exercise are:

1. to describe the baseline characteristics of the sample of patients included;
2. to demonstrate that the randomization has worked well by achieving well balanced treatment groups at baseline;
3. to add credibility to the trial results, specifically encouraging confidence in unadjusted outcome analyses as being without any serious bias;
4. to identify any unlucky imbalances between treatment groups that may have arisen by chance.

The first of these aims is perhaps the most useful, since it is important to document who is in the trial so that we can assess to whom the trial findings can be extrapolated. However, it does not actually require treatment comparison as such.

Most trials have a 'Table I', for example, 46 (92 per cent) did so in our survey, and it is worth dwelling on their content, as shown in Table III. The reports vary enormously in the number of baseline features included, with a median of 14 and a maximum of 41, the latter [22] being a particularly enormous table occupying nearly a whole journal column. It might be helpful if some authors were more parsimonious in their inclusion of variables in 'Table I', focusing on a smaller subset of key variables that either crucially define the patient sample and/or are likely predictors of patient outcome. Perhaps authors could have available on request a more extensive list of baseline features, and of course more expansive reports, such as for regulatory submissions, can include a larger baseline table.

One stylistic excess is to present the baseline results both for each treatment group and for all groups combined, whereas one or other would be sufficient. A further stylistic issue is the duplication inherent in giving both baseline counts and percentages. Again, for such background data, perhaps one or the other is sufficient. Also, the inappropriate use of standard errors for baseline variables was mostly avoided, the more informative standard deviations being better descriptors of between-patient variation.

Another contentious issue is the use of significance tests for baseline comparison. It is a common practice; 24 trials (48 per cent) in our survey performed such tests. As far as we

Table III. Baseline variables by treatment group in 50 clinical trial reports.

		Number of trials
Number of baseline variables compared	None	4
	One to four	1
	Five to nine	14
	Ten to 19	24
	20 to 29	5
	30 or more	2
Significance tests for baseline difference performed	Yes	24
	No	26
Baseline imbalances noted	Yes	17
	No	33

can calculate, 299 tests were thus performed of which 18 (6 per cent) reached $p < 0.05$. By definition, all baseline differences are due to chance (unless the randomization goes wrong).

As illustrated in Section 3, the importance of a baseline difference depends on the variable's strength of association with outcome and the standardized magnitude of this difference, called Z_x. A strong predictor's imbalance could matter without being statistically significant, while such statistical significance is irrelevant for a baseline variable not related to outcome. Thus, P-values for baseline differences do not serve a useful purpose [23, 24], since they are not testing a useful scientific hypothesis.

Particularly excessive is the addition of an extra column of P-values in the table of baseline data [25]. Others report the significance of baseline differences, but give no insight into whether such factors were related to outcome [26]. Thus, of the 17 trials in which baseline differences were noted, most ignored this fact when analysing the outcome results.

5. CONCLUSIONS

While there have been improvements in the standards of reporting for clinical trials, aided by guidelines such as CONSORT [27, 28], this paper has identified some important deficiencies and inconsistencies as regards the uses of baseline data in clinical trial reports.

Subgroup analyses are often given too great a prominence and fail to use appropriate methods of statistical inference such as interaction tests. There also appears a lack of consistency regarding the use of covariate-adjusted analyses, perhaps largely because their rationale and statistical properties are poorly understood. Though less serious, there are also improvements to be made in the reporting of baseline comparisons.

By linking our methodological arguments to a survey of recent trial reports, we intended to convey a sense of reality to the discussion on how to best undertake and report subgroup analyses and covariate adjustments. We hope this approach enhances the debate and the development of clearer guidelines for statistical reporting.

Of overriding importance is the need for each clinical trial to define a clear and coherent policy for such uses of baseline data in the context of an overall predefined statistical analysis plan. Thus, the risk of *post hoc* exaggerated emphases across a multiplicity of possible analyses can be reduced, and readers can have greater confidence in the validity of authors' conclusions.

While subgroup analysis may have the more obvious potential for 'statistical sins', there is more need for statistical debate and methodological research on what truly constitutes the best strategy for covariate-adjusted analyses. We have tried to give some useful insights and recommendations, but still feel that the statistical properties of covariate adjustment, in the common situation when one does not know the important outcome predictors in advance, still need to be more fully understood.

REFERENCES

1. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; **355**:1064–1069.
2. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. *Journal of the American Medical Association* 1991; **266**:93–98.
3. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *New England Journal of Medicine* 1987; **317**:426–432.
4. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
5. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**:871–882.
6. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statistics in Medicine* (in press).
7. Frasure-Smith N, Lespérance F, Prince RH, Verrier P, Garber R, Juneau M, Wolfson C, Bourassa M. Randomised trial of home-based psychological nursing intervention for patients recovering from myocardial infarction. *Lancet* 1997; **350**:473–479.
8. Kostis JB, David BR, Cutler J, Grimm RH, Berge KG, Cohen JD, Lacy CR, Perry HM, Blaufox MD, Wassertheil-Smoller S, Black HR, Schron E, Berkson DM, Curb JD, McFate Smith W, McDonald R, Applegate WB. Prevention of heart failure by antihypertensive drug treatment in older persons with isolated systolic hypertension. *Journal of the American Medical Association* 1997; **278**:212–216.
9. Altman DG. Adjustment for covariate imbalance. In *Encyclopaedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: Chichester, 1998; 1000–1005.
10. Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. *Controlled Clinical Trials* 2000; **21**:330–342.
11. Senn SJ. Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* 1989; **8**:467–475.
12. Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; **2**:1685–1704.
13. Snedecor GW, Cochran WG. Analysis of covariance. In *Statistical Methods*. Iowa State University Press: Ames, 1989; 419–446.
14. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **59**:227–240.
15. Ford I, Norrie J, Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* 1995; **14**:735–746.
16. Ford I, Norrie J. The role of covariates in estimating treatment effects and risk in long-term clinical trials. *Statistics in Medicine*.
17. Schluchter MD, Forsythe AB. Post-hoc selection of covariates in randomised experiments. *Communications in Statistics — Theory and Methods* 1985; **14**:679–699.
18. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials* 1989; **10**:161S–175S.
19. Van der Horst CM, Saag MS, Cloud GA, Hamill RJ, Graybill JR, Sobel JD, Johnson PC, Tuazon CU, Kerkering T, Moskovitz BL, Powderly WG, Dismukes WE. The National Institute of Allergy and Infectious Diseases Mycoses Study Group and AIDS Clinical Trials Group. Treatment of cryptococcal meningitis associated with the acquired immunodeficiency syndrome. *New England Journal of Medicine* 1997; **337**:15–21.
20. Olds DL, Eckenrode J, Henderson CR, Kitzman H, Powers J, Cole R, Sidora K, Morris P, Pettitt LM, Luckey D. Long-term effects of home visitation on maternal life course and child abuse and neglect. *Journal of the American Medical Association* 1997; **278**:637–643.

21. Rolfs RT, Riduan Joesoef M, Hendershot EF, Rompalo AM, Augenbraun MH, Chiu M, Bolan G, Johnson SC, French P, Steen E, Radolf JD, Larsen S. A randomised trial of enhanced therapy for early syphilis in patients with and without human immunodeficiency virus infection. *New England Journal of Medicine* 1997; **337**:307–314.
22. Sundberg K, Bank J, Smidt-Jensen S, Brocks V, Lundsteen C, Parner J, Keiding N, Philip J. Randomised study of risk of fetal loss related to early amniocentesis versus chorionic villus sampling. *Lancet* 1997; **350**:697–703.
23. Senn S. Testing for baseline balance in clinical trials. *Statistics in Medicine* 1994; **13**:1715–1726.
24. Altman DG. Comparability of randomised groups. *Statistician* 1985; **34**:125–136.
25. Gordin FM, Matts JP, Miller C, Brown LS, Hafner R, John SL, Klein M, Vaughn A, Besch CL, Perez G, Szabo S, El-Sadr W. A controlled trial of isoniazid in persons with energy and human immunodeficiency virus infection who are at high risk for tuberculosis. *New England Journal of Medicine* 1997; **337**:315–320.
26. Columbus Investigators. Low-molecular-weight heparin in the treatment of patients with venous thromboembolism. *New England Journal of Medicine* 1997; **337**:657–662.
27. 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**:1191–1194.
28. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T. The revised CONSORT statement for reporting randomised trials: explanation and elaboration. *Annals of Internal Medicine* 2001; **134**:663–694.