Epidemiology 5

Multiplicity in randomised trials II: subgroup and interim analyses

Kenneth F Schulz, David A Grimes

Lancet 2005; 365: 1657–61

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

Subgroup analyses can pose serious multiplicity concerns. By testing enough subgroups, a false-positive result will probably emerge by chance alone. Investigators might undertake many analyses but only report the significant effects, distorting the medical literature. In general, we discourage subgroup analyses. However, if they are necessary, researchers should do statistical tests of interaction, rather than analyse every separate subgroup. Investigators cannot avoid interim analyses when data monitoring is indicated. However, repeatedly testing at every interim raises multiplicity concerns, and not accounting for multiplicity escalates the false-positive error. Statistical stopping methods must be used. The O'Brien-Fleming and Peto group sequential stopping methods are easily implemented and preserve the intended α level and power. Both adopt stringent criteria (low nominal p values) during the interim analyses. Implementing a trial under these stopping rules resembles a conventional trial, with the exception that it can be terminated early should a treatment prove greatly superior. Investigators and readers, however, need to grasp that the estimated treatment effects are prone to exaggeration, a random high, with early stopping.

Subgroup analyses have specious appeal. They seem logical and intuitive—and even fun—to both investigators and readers. However, this insidious appeal causes important problems. Multiplicity and naiveté combine to encourage interpretational missteps in trial conduct and reporting. The subgroup treatment effects revealed in many reports might be illusory.

By contrast, investigators cannot avoid interim analyses if data monitoring is indicated. Neither can they use their normal statistical approaches at interim analyses. Statistical stopping methods, essentially statistical adjustments for warning rather than stopping, must be used in support of data monitoring. Unfortunately, those methods baffle investigators and readers alike. Statistics frequently proves confusing anyway without throwing in second-order complications of stopping methods.

Multiplicity issues from subgroup and interim analyses pose similar problems to those from multiple endpoints and treatment groups.[1] Investigators frequently data-dredge by doing many subgroup analyses and undertaking repeated interim analyses. Also, researchers conduct unplanned subgroup and interim analyses. Yet, some of the approaches to multiplicity problems from subgroup and interim analyses differ from those for endpoints and treatments.

Subgroup analyses

Indiscriminate subgroup analyses pose serious multiplicity concerns. Problems reverberate throughout the medical literature. Even after many warnings,[2] some investigators doggedly persist in undertaking excessive subgroup analyses.

Investigators define subgroups of participants by characteristics at baseline. They then do analyses to assess whether treatment effects differ in these subgroups. The major problems stem from investigators undertaking statistical tests within every subgroup examined. Combining analyses of multiple subgroups with multiple outcomes leads to a profusion of statistical tests.

Seeking positive subgroup effects (data-dredging), in the absence of overall effects, could fuel much of this activity. If enough subgroups are tested, false-positive results will arise by chance alone.

> "The answer to a randomized controlled trial that does not confirm one's beliefs is not the conduct of several subanalyses until one can see what one believes. Rather, the answer is to re-examine one's beliefs carefully."[3]

Similarly, in a trial with a clear overall effect, subgroup testing can produce false-negative results due to chance and lack of power.

The Lancet published an illustrative example.[4] Aspirin displayed a strongly beneficial effect in preventing death after myocardial infarction ($p<0\cdot00001$, with a narrow confidence interval). The editors urged the researchers to include nearly 40 subgroup analyses.[2] The investigators reluctantly agreed under the condition that they could provide a subgroup analysis of their own to illustrate their unreliability. They showed that participants born under the astrological signs Gemini or Libra had a slightly adverse effect on death from aspirin (9% increase, SD 13; not significant) whereas participants born under all other astrological signs reaped a strikingly beneficial effect (28% reduction, SD 5; $p<0\cdot00001$).[4]

See *Lancet* 2005; **365:** 1591–95

Anecdotal reports of support from astrologers to the contrary, this chance zodiac finding has generated little interest from the medical community. The authors concluded from their subgroup analyses that:

> "All these subgroup analyses should, perhaps, be taken less as evidence about who benefits than as evidence that such analyses are potentially misleading."

These and other thoughtful investigators stress that usually the most reliable estimate of effect for a particular subgroup is the overall effect (essentially all the subgroups combined) rather than the observed effect in that subgroup.[4,5] We agree.

Proper analysis dissipates much of the multiplicity problem with subgroup analyses. Frequently, investigators improperly test every subgroup, which opens the door to chance findings. For example, breaking down age at baseline into four categories yields four tests just on that characteristic (table 1). A proper analysis uses a statistical test of interaction, which involves assessing whether the treatment effect on an outcome depends on the participant's subgroup. That not only tests the proper question but also produces a single test instead of four, substantially addressing the multiplicity problem. Investigators have questioned interaction tests based on lack of power. However, interaction tests provide proper caution. They recognise the limited information available in the subgroups and have emerged as the most effective statistical method to restrain inappropriate subgroup findings, while still having the ability to detect interactive effects, if present.[6,7]

	Febrile morbidity			Rate ratio (95% CI)
	Yes	No	Total	
Age 20–24 years				
New antibiotic	11	84	95	1·4 (0·6–3·2)
Standard antibiotic	8	86	94	
Age 25–29 years				
New antibiotic	8	69	77	1·2 (0·4–3·1)
Standard antibiotic	7	72	79	
Age 30–34 years				
New antibiotic	3	48	51	0·3 (0·1–0·9)
Standard antibiotic	11	38	49	
Age 35–39 years				
New antibiotic	10	32	42	1·1 (0·5–2·5)
Standard antibiotic	9	33	42	
Total				
New antibiotic	32	233	265	0·9 (0·6–1·4)
Standard antibiotic	35	229	264	

The test for statistical interaction (Breslow-Day) is non-significant (p=0·103), suggesting that a subgroup finding in the 30–34 age stratum is attributable to chance. However, that result, if inappropriately highlighted, would be an example of a superfluous subgroup salvage of an otherwise indeterminate (negative) trial.

***Table 1:* Effect of new versus standard antibiotic on febrile morbidity in four age strata and overall**

Another problem with subgroup analyses is that investigators can do many analyses and only report the significant ones, which bestows more credibility on them than they deserve—a misleading practice and, if intentional, unethical. This situation is analogous to what we judge a major problem with multiple endpoints.

Subgroup analyses remain a problem in published work. In a review of 50 reports from general medical journals (*New England Journal of Medicine*, *The Lancet*, *JAMA*, and *BMJ*), 70% reported subgroup analyses.[8] Of those in which the number of analyses could be established, almost 40% did at least six subgroup analyses—one reported 24. Fewer than half used statistical tests of interaction. Furthermore, the reports did not provide information on whether the subgroup analyses were predefined or post hoc. The authors of the review suspected 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".[8] Disappointingly, most trials reporting subgroup analyses noted a subgroup difference that was highlighted in the conclusions[8]—so much for cautious interpretation!

We discourage subgroup analyses. If properly undertaken they are not necessarily wrong. Sometimes they make biological sense or they are mandated by sponsors, both public and industry. If done, they should be confined to the primary outcome and a limited number of subgroups. Those planned should be prespecified in the protocol. Investigators must report all subgroup analyses done, not just the significant ones. Importantly, they should use statistical tests of interaction to assess whether a treatment effect differs among subgroups rather than individual tests within each subgroup. This approach alleviates major concerns with multiple comparisons. Rarely should subgroup analyses affect the trial's conclusions.

> "Subgroup analyses are particularly prone to over interpretation, 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."[8,9]

Methodologists have been too restrained in criticising improperly undertaken subgroup analyses. Stronger denunciation is needed.

What readers should look for with subgroup analyses

Readers should be wary of trials that report many subgroup analyses, unless the investigators provide valid reasons. Also, beware of trials that provide a small number of subgroup analyses. They might have done many and just cherry-picked the interesting and significant ones. Consequently, faulty reporting could mean that trials with few subgroup analyses are even worse than the trials with many. Investigators have more credence if they state that they reported all the analyses done. Furthermore, researchers should label non-prespecified subgroup analyses as hypothesis-generating rather than confirming. Such findings should not appear in the conclusions.

Readers should expect interaction tests for subgroup effects. Discount analyses built on tests within subgroups. Even with a significant interaction test, readers should base interpretation of the findings on biological plausibility, on prespecification of analyses, and on the statistical strength of the information. Generally, adjustments for multiplicity are unnecessary when investigators use interaction tests. However, in view of the frequently frivolous data-dredging pursuits involved, the argument for statistical adjustments is stronger than that for multiple endpoints. Moreover, if investigators do not use interaction tests and report tests on every individual subgroup, multiplicity adjustments are appropriate.[10] Most subgroup findings tend to exaggerate reality. Be especially suspicious of investigators highlighting a subgroup treatment effect in a trial with no overall treatment effect.[11] They are usually superfluous subgroup salvages of otherwise indeterminate (negative) trials (table 1).[8]

Interim analyses

Appropriate monitoring of trials involves more than statistical warnings for stopping. Indeed, the superiority or inferiority of the studied treatment has a major role. However, slow accrual, poor data quality, poor adherence, resource deficiencies, unacceptable adverse effects, fraud, and emerging information that make the trial irrelevant, unnecessary, or unethical, all could lead to stopping a trial. The decision process is clearly complex.[12,13] It best resides with an independent data monitoring committee. The committee's task becomes manageable with a prespecified statistical stopping method. Yet, investigators and readers frequently remain oblivious to these statistical issues.

Accumulating data in trials tempt investigators to do analyses on the main endpoint. If they seek $p<0\cdot05$ at the end of the study, they might still undertake all the interim analyses at $\alpha=0\cdot05$. That is wrong.

A graphical depiction of an example perhaps clarifies the issue (figure). A data monitoring committee does an interim analysis every 6 months for 5 years. At 18 months, the analysis slips under $p<0\cdot05$, but never again attains significance at that level. An early decision by the committee to stop the trial based on this result might have led to an incorrect conclusion about the effectiveness of the intervention.

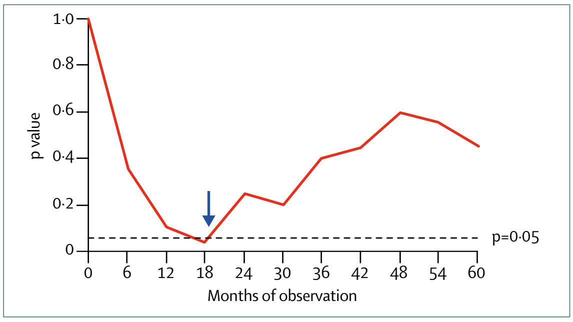

***Figure:* Interim analyses done every 6 months for 5 years**
The p value is shown for the comparison between the treatment group and control group.

Intuitively, undertaking many interim analyses at $p<0\cdot05$ should actually inflate the false-positive-error rate (α). Indeed, if an investigator looks at the accumulating data at $\alpha=0\cdot05$ at every interim, then the actual overall α level rises with the number of challenges—eg, overall $\alpha=0\cdot08$ after two challenges, $\alpha=0\cdot11$ after three, and $\alpha=0\cdot19$ after ten.[9,13] This multiplicity problem dictates the need for statistical adjustment: scientific credibility depends on it.

Methodologists have developed many statistical stopping (actually, warning) procedures, sometimes called data-dependent stopping rules or guidelines.[13] If investigators undertake interim analyses, they must use one of these procedures. The group sequential designs have garnered perhaps the most attention. They tend to be easier to understand, construct, and apply.[14] On the basis of the number of interim analyses planned, the methods define p values for considering trial stoppage at an interim look while preserving the overall type I error (α; table 2).

Number of planned interim analyses	Interim analysis	Pocock	Peto	O'Brien-Fleming
2	1	0·029	0·001	0·005
	2 (final)	0·029	0·05	0·048
3	1	0·022	0·001	0·0005
	2	0·022	0·001	0·014
	3 (final)	0·022	0·05	0·045
4	1	0·018	0·001	0·0001
	2	0·018	0·001	0·004
	3	0·018	0·001	0·019
	4 (final)	0·018	0·05	0·043
5	1	0·016	0·001	0·00001
	2	0·016	0·001	0·0013
	3	0·016	0·001	0·008
	4	0·016	0·001	0·023
	5 (final)	0·016	0·05	0·041

Overall $\alpha=0\cdot05$.

***Table 2:* Interim stopping levels (p values) for different numbers of planned interim analyses by group sequential design[14,15]**

The fixed nominal level approach (Pocock approach) proves simple and allows fairly early termination of trials. However, it suffers from the final test of significance being at a smaller p value than that of a regular fixed-sample trial. For example, to yield an overall α=0·05 with three interim analyses, investigators would have to test at 0·022 at each analysis, including the final one (table 2). If the final test yielded p=0·03, then the trial would be deemed not significant by this group sequential approach, but it would have been significant if the group sequential approach had not been used. This approach is mainly of historical interest because other methods incorporate its advantages without this disadvantage.[14]

We favour two other procedures: O'Brien-Fleming and Peto.[12–14] Both adopt stringent criteria (low nominal p values) during the interim analyses (table 2). If the trial continues until the planned sample size, then all analyses proceed as if basically no interim analyses had taken place. The procedures preserve not only the intended α level but also the power.[16] Data are obtained in essentially the same way as in a fixed-sample design. Their beauty is in simplicity. Implementation of a trial under these stopping rules mirrors that of a conventional trial, with the exception that the trial can be terminated early should a treatment prove greatly superior. As a general rule, investigators gain little by doing more than four or five interim analyses during a trial.[9,17] Thus, with minimal additional effort, researchers address the ethical need to monitor for substantial treatment effects, positive or negative.

The Peto (or Haybittle-Peto) approach is simpler to understand, implement, and describe. It uses constant but stringent stopping levels until the final analysis (table 2). For some trials, however, investigators believe that early termination of a trial is too difficult with Peto.

The O'Brien-Fleming approach appeals intuitively to many investigators because the stopping criteria are conservative early on, when everyone should be dubious of unstable results, and they successively ease as the results become more reliable and stable. Unlike the Peto approach, the O'Brien-Fleming stopping criteria vary with every interval look at the data.

See *Lancet* 2005; **365:** 1348–53

If investigators plan interim analyses, they should prespecify the statistical stopping approach. Furthermore, an independent trial statistician, rather than the researchers, should do the analyses for the data monitoring committee.[13] The interim analysis plan could be in the protocol, in a separate statistical analysis plan, or in a data monitoring committee charter. The analysis plan and charter, if appropriate, can be appendices to the protocol. Having them as appendices keeps the protocol more approachable to the implementation staff undertaking the trial.[13]

Most trials probably do not need an interim analysis and independent monitoring.[18] Of 662 eligible trials identified in 2000, 24% mentioned use of a data monitoring committee, interim analyses, or both.[19]

Early termination and biased estimates of treatment effects

If a data monitoring committee stops a trial early on the basis of a group sequential stopping procedure, the estimates of treatment effect are biased. That remains a shortcoming of these procedures. As explanation, suppose the investigators did the same trial many times. Random fluctuations towards greater treatment effects would more probably result in early termination than random fluctuations towards lesser treatment effects. Thus, when a trial is stopped early, readers need to grasp that the estimated treatment effects are prone to exaggeration—ie, a random high.[12,14] When an unbiased estimate is paramount, investigators should focus on a fixed sample design and shun group sequential designs.

Stopping for harm or futility

Thus far in our discussion of stopping guidelines, we have implied the same level of evidence to terminate early irrespective of whether for benefit or harm. Methodologists call such a strategy symmetric stopping boundaries with group sequential methods, analogous to two-sided hypothesis testing.

Some investigators or data monitoring committees, however, might desire asymmetric stopping boundaries. These allow for a lower level of evidence to terminate for harm than for benefit. For example, O'Brien-Fleming sequential boundaries might be used in monitoring for benefit whereas Pocock-type sequential boundaries could be used when monitoring for harm.[13]

Sometimes researchers or a data monitoring committee do not want to establish harm. Alternatively, they desire to denote trends that are sufficiently unfavourable, such that completion of the trial is unlikely to yield a significant beneficial effect. That facilitates stopping for futility, which only permits an assertion of inability to establish benefit. This approach engenders fancy terminology: conditional power or stochastic curtailment.

With conditional power, investigators design trials with a stated power.[20] However, once investigators start the trial, sustained data accumulation enriches knowledge (shielded from the investigators, of course). With accumulating data, the power can be recalculated. For example, an emerging trend towards treatment efficacy increases power whereas an unfavourable trend reduces power. Conditional power describes this evolving power estimate.

Monitoring groups use conditional power most frequently with trends unfavourable to treatment. If the conditional power calculations yield low power for various assumed treatment effects, including the assumed treatment effect in the trial protocol, then a monitoring group might consider continuation of the trial as futile and recommend stopping. This implementation of conditional power breeds the terms stochastic curtailment and stopping for futility. These methods have been used effectively in monitoring trials.[13,14]

Other statistical stopping methods

Other statistical stopping methods also have appeal. Lan-DeMets (alpha spending function) developed a more flexible adaptation of group sequential methods.[21,22] It controls the false-positive error used at every interim analysis as a function of the proportion of total information observed, which allows the number and exact timing of the interim analyses to change after the trial has started.[13,14] The data monitoring committee begins with a schedule that could change on the basis of emerging data. Thus, an alpha spending approach allows for unplanned looks.

We find Bayesian approaches helpful in clinical decision making,[23] but remain sceptical about their use in interim analyses. Bayesian approaches represent a separate branch of statistics. Correctly implemented, they can be useful for monitoring.[24–27] Readers, however, have little need to understand them, because they are used sparingly. Moreover, they elicit concerns. For example, every interim challenge might be at the 0·05 level, seriously escalating the overall false-positive error rate (α).[13] Unfortunately, some sponsors might take a keen interest in this higher likelihood of finding a significant effect.

See *Lancet* 2005; **365:** 1500–05

What readers should look for with interim analyses

Readers should remain alert for unreported interim analyses. If they find a statement from the researchers that no interim analyses were done, multiplicity is probably not a problem. However, such transparent reporting rarely happens. Poor reporting might camouflage the interim looks that the investigators did. Admittedly, detection of such interim analyses poses problems for readers. One clue is that the calculated p value is slightly less than 0·05, which could mean the researchers repeatedly tested and stopped the trial just when $p<0·05$ was attained. Another clue might be if the completed trial size is less than that planned. One reason that sample size calculations are desired in the methods section is to indicate if the trial stopped early. Readers should be wary if the trial stopped early and no statistical stopping rule was described.

If researchers describe a statistical stopping rule, readers should evaluate its appropriateness. Peto and O'Brien-Fleming methods accomplish the goals of interim analyses without detracting from the trial. The other statistical approaches to interim analyses, most of which sport fancy names like alpha spending function and conditional power, usually are appropriate, but Bayesian approaches can provoke concerns.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

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

References

1 Schulz KF, Grimes DA. Multiplicity in randomised trials I: endpoints and treatments. *Lancet* 2005; **365:** 1591–95.

2 Horton R. From star signs to trial guidelines. *Lancet* 2000; **355:** 1033–34.

3 Oei SG, Helmerhorst FM, Keirse MNC. Postcoital test should be performed as routine infertility test. *BMJ* 1999; **318:** 1008–09.

4 ISIS-2 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.

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

6 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002; **21:** 2917–30.

7 Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ* 2003; **326:** 219.

8 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000; **355:** 1064–69.

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

10 Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998; **316:** 1236–38.

11 Dyson DC, Crites YM, Ray DA, Armstrong MA. Prevention of preterm birth in high-risk patients: the role of education and provider contact versus home uterine monitoring. *Am J Obstet Gynecol* 1991; **164:** 756–62.

12 Pocock SJ. When to stop a clinical trial. *BMJ* 1992; **305:** 235–40.

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

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

15 Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* 1987; **43:** 213–23.

16 O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35:** 549–56.

17 McPherson K. Sequential stopping rules in clinical trials. *Stat Med* 1990; **9:** 595–600.

18 Sydes MR, Spiegelhalter DJ, Altman DG, Babiker AB, Parmar MKB, and the DAMOCLES Group. Systematic qualitative review of the literature on data monitoring committees for randomized controlled trials. *Clin Trials* 2004; **1:** 60–79.

19 Sydes MR, Altman DG, Babiker AB, Parmar MKB, Spiegelhalter D, and the DAMOCLES group. Reported use of data monitoring committees in the main published reports of randomized trials: a cross-sectional study. *Clin Trials* 2004; **1:** 48–59.

20 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; **365:** 1348–53.

21 Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70:** 659–63.

22 DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. *Stat Med* 1994; **13:** 1341–52.

23 Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet* 2005; **365:** 1500–05.

24 Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. *Stat Med* 1993; **12:** 1501–11.

25 Freedman LS, Spiegelhalter DJ, Parmar MK. The what, why and how of Bayesian clinical trials monitoring. *Stat Med* 1994; **13:** 1371–83.

26 Parmar MK, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. *Stat Med* 1994; **13:** 1297–312.

27 Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E. Monitoring of large randomised clinical trials: a new approach with Bayesian methods. *Lancet* 2001; **358:** 375–81.