How to interpret figures in reports of clinical trials

A picture may be worth a thousand words but in medical research, caution **Stuart Pocock**, **Thomas Travison**, and **Lisa Wruck**, it is important to understand exactly what you are looking at

Stuart J Pocock professor of medical statistics, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, London WC1E 7HT

Thomas G Travison senior research scientist, New England Research Institutes, Watertown, MA, USA

Lisa M Wruck senior biostatistician, Rho Inc, Chapel Hill, NC, USA

Correspondence to: stuart.pocock@lshtm.ac.uk

Accepted: 15 February 2008

The graphical display of data is among the most powerful tools available for communicating medical research findings, given the increasing complexity of study designs and the mind's preference for information conveyed in pictorial format.[1 2] However, although general information is available on what constitutes an effective data display[1-6] and what constitutes good practice in reporting trials,[7 8] there is relatively little guidance on using figures to aid the presentation of trial results.[9]

Because figures are so effective in creating an enduring impression of results, their construction—and interpretation by readers—must be handled with care. We recently conducted a survey to determine the types of figures used most commonly in reports of clinical trials and to uncover the good, and not so good, practices that typically attend their use.[10] Here, we highlight the important features of the most commonly used types of figures. In doing so, we hope to illustrate the hallmarks of figures that are likely to convey an impression consistent with valid trial conclusions and those aspects of figures that may, without careful interpretation, be misleading.

What comes up most

We examined all issues of five major general medical journals (*Annals of Internal Medicine, BMJ, JAMA, Lancet,* and *New England Journal of Medicine*) published from November 2006 to January 2007. The 77 reports of randomised trials included in these journal issues contained 175 figures (mean 2.3 figures per article). The four most common types of figure were flow diagrams (66 articles), Kaplan-Meier plots (32 articles), forest plots (21 articles), and repeated measures plots (20 articles) (see table on bmj.com).[10]

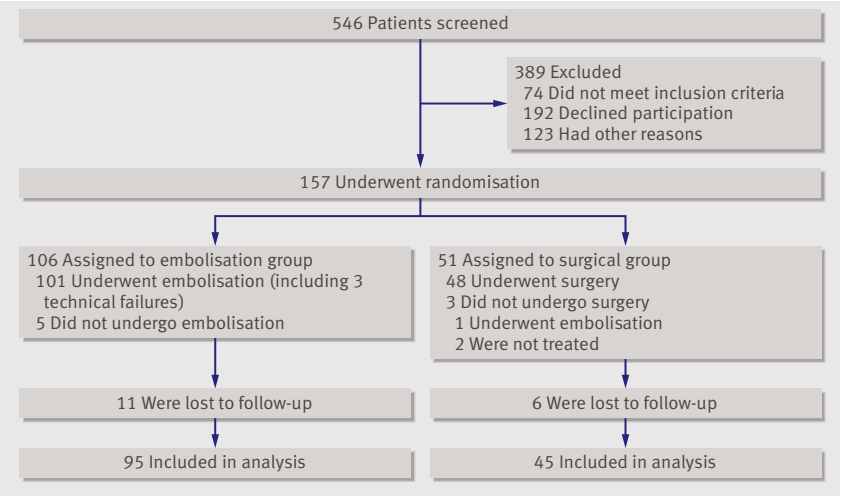


Fig 1 | Flow diagram from study comparing uterine-artery embolisation with surgery for fibroids (redrawn from Edwards et al[11])

Flow diagrams

Flow diagrams are integral to the CONSORT guidelines for the reporting of clinical trials.[7 8] They display the flow of participants through the stages of the trial in a way that should be easy to follow. Figure 1 depicts a successful example of a flow diagram portraying a clear picture of the trial's design and conduct. It includes the numbers of people screened and reasons for exclusion, information that many trials fail to collect and report. The numbers not receiving randomised treatment and numbers lost to follow-up are key limitations that every study should document.

The flow chart in fig 2 is harder to read because it contains substantial repetition of words. Such flow diagrams in multi-arm trials may be more concisely displayed as a table, provided there is no loss of information.

Kaplan-Meier plots

The Kaplan-Meier plot is for time to event or survival data, when interest is focused on the risk of a particular event (such as death or myocardial infarction) as participants move through time.[13] Because the aim of many treatments or interventions is to try to reduce the occurrence of a particular event, this type of plot is used commonly in reporting clinical trials. However, it is an aspect of statistics not well understood by doctors.[14]

The plot is drawn with time in the study on the horizontal axis and either the cumulative proportion with the event, or the proportion for whom the event has not yet occurred (the survival probability), plotted on the vertical. Curves are drawn for each treatment group, and the separation between the curves indicates potential differences in the treatments' effectiveness. The Kaplan-Meier estimates change only when events actually occur, so that each plot is a series of steps. Note how few participants were followed to five years.

Figure 3 shows the essential features of a clear Kaplan Meier plot. The treatment groups are visually differentiable, with an appropriate vertical scale and axes clearly labelled. Below the horizontal axis, the numbers of participants remaining at risk (that is, those who remain under observation and for whom the event is yet to occur) are displayed.

A formal statistical comparison (in this case a hazard ratio with 95% confidence interval and P value from the logrank test) is needed to assess whether the distance between the curves is sufficient to depict a real difference in risk between treatment and control arms. This information is often best included on the figure itself. In this case the slight difference is not significant.

Figure 4 shows a plot going down (plotting the proportion of participants who are event free), covering

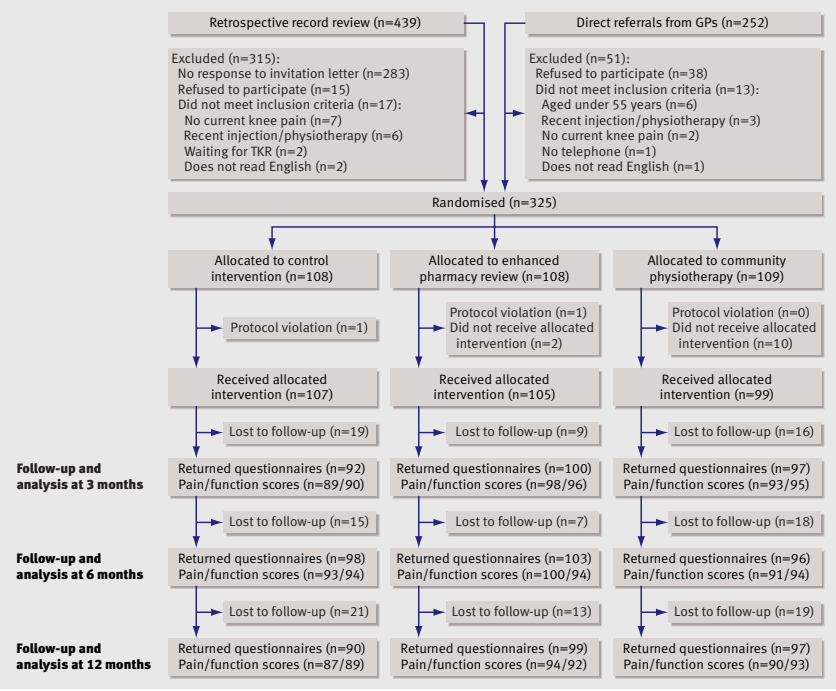


Fig 2 | Flow diagram for randomised trial of three primary care strategies for knee pain[12]

the whole scale from probability 1 to 0. Much of the graph is empty space because the event (defaulting from treatment) has low incidence. For outcomes of this type, it is more useful to present the cumulative probability curve going up, with the vertical axis truncated at a reasonable maximum.

It can be helpful if Kaplan-Meier plots take account of statistical uncertainty by displaying standard error bars (or confidence intervals) at a few key follow-up times[13] to help restrain readers from overinterpreting any apparent differences between the curves.

Forest plots

A forest plot displays estimated treatment effects across various patient subgroups.[17] Typically, a forest plot presents an overall effect (for all randomised participants) and then various subgroup computations (for instance, by sex) on a common axis. Each point plotted represents a comparison between treatment and control participants in the relevant subgroup and is accompanied by its 95% CI.

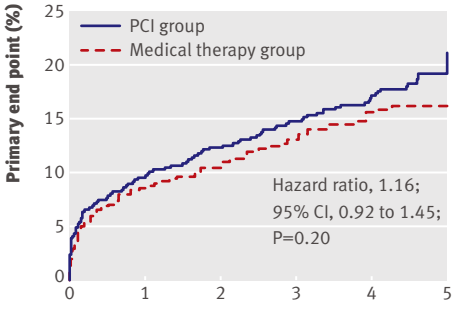
Figure 5 is a simple example of a forest plot, with only one set of subgroup analyses.[18] This figure has several features consistent with good practice. It shows the overall estimate and confidence intervals (combining all subgroups) and the labels indicate which direction favours treatment or control. Subgroup estimates are displayed underneath the overall estimate. Although the lines suggest that patients with a baseline albumin concentration below 25 g/l may benefit from albumin treatment, inclusion of the heterogeneity test (sometimes called interaction test) makes it clear that the evidence is not strong enough to be conclusive. Such interaction tests are key to interpretation of forest plots[19 20] and should be included on the plot or in the legend.

When forest plots display ratios (as in fig 5) rather than absolute differences, the horizontal axis may be on a logarithmic scale, so that a ratio of 2 is depicted as being as far away from 1 as is 0.5. This makes sense because 2 is the multiplicative inverse of 1/2.

Extra numerical information is often tabulated alongside the figure. In fig 5 the numbers of deaths and patients (to the left of the plotted estimates) are helpful as they are the "raw data" for each subgroup.

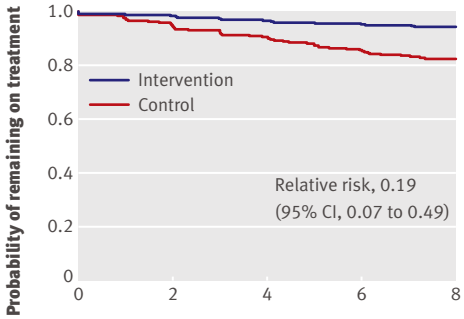
Most forest plots present several subgroup analyses (fig 6). Presentation of results in both tabular and graphical format allows readers to examine the effects with precision and facilitates inclusion of data in subsequent meta-analyses. However, fig 6 does not give the results of heterogeneity tests; instead the authors state "there was no evidence of substantial heterogeneity."

This figure displays some additional conventions consistent with good practice. Vertical lines are plotted both for the value indicating no treatment effect (dotted at 1.0) and for the overall effect (solid, at 0.64). In addition, the size of the plotted symbol for point estimates is proportional to the number of events within each subgroup. Forest plots are also used in meta-analyses



No at risk						
PCI group	1082	895	719	482	265	85
Medical therapy group	1084	909	714	474	268	78

Fig 3 | Kaplan-Meier curve from trial of percutaneous coronary intervention (PCI) for persistent occlusion after myocardial infarction. The primary end point was death from any cause, non-fatal reinfarction, or heart failure requiring hospital admission (Redrawn from Hochman et al[15])



No at risk					
Intervention	778	744	719	694	497
Control	744	683	635	590	392

Fig 4 | Kaplan-Meier plot from trial of strategies to improve adherence to tuberculosis treatment (redrawn from Thiam et al[16])

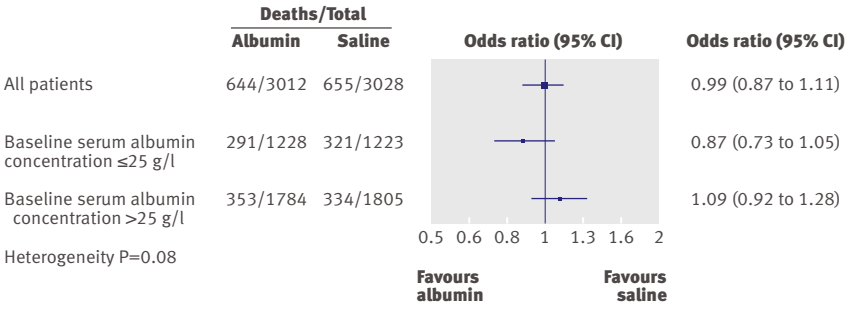


Fig 5 | Forest plot from study comparing resuscitation with albumin or saline in intensive care showing unadjusted odds ratio of death stratified by baseline albumin concentration[18]

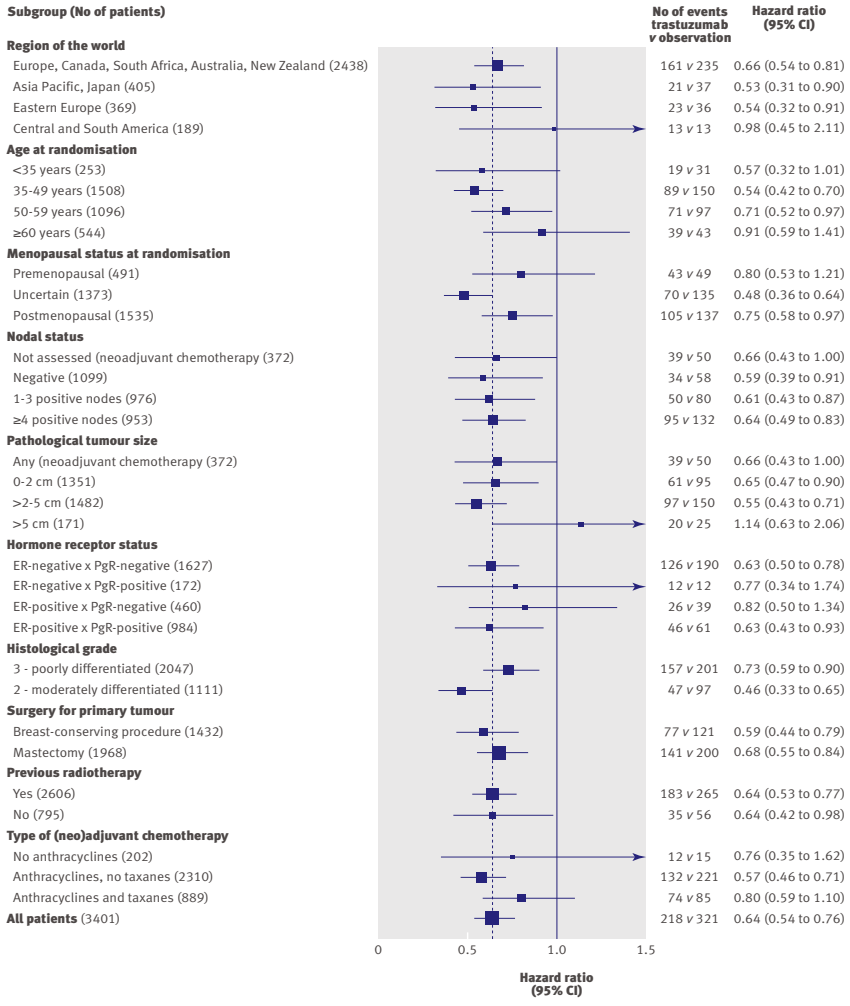


Fig 6 | Forest plot from trial of adjuvant chemotherapy in HER2-positive breast cancer showing hazard ratio for disease free survival (redrawn from Smith et al[21])

combining evidence from several related studies, where the same issues arise.

Repeated measures plots

For trials with a quantitative outcome measured at baseline and two or more follow-up times, it is common to plot the means by treatment over time. Figure 7 shows this approach in a clear style for three outcome scores each recorded at baseline and five follow-up times. The figure uses different symbols for each treatment to help distinguish them, and joining the means by lines helps the eye to follow the trends over time.

As with the forest plots, it is important to express the statistical uncertainty in each mean; this is done here using confidence intervals. To enhance clarity, the authors have helpfully staggered group means at each time to ensure that intervals do not obscure each other. The figure also includes a global P value corresponding to a test of overall differences in outcomes between study arms. This avoids the undesirable use of repeated significance tests at every time point and the consequent problem of inflated type I error due to multiple testing.

As a longitudinal study will typically lose some participants with time, it would have been useful to give the numbers of participants at each time point under the x axis, as in fig 3.

Figure 8 uses a different approach to displaying longitudinal trends, plotting mean changes from baseline rather than means. Analyses of covariance adjusting for baseline value is a preferred method of inference for such data.[23] The numbers of patients by group at each time are given below the x axis. The authors documented the statistical comparison of treatments at final visit, an important detail that clarifies that the observed treatment differences remained significant. However, their use of

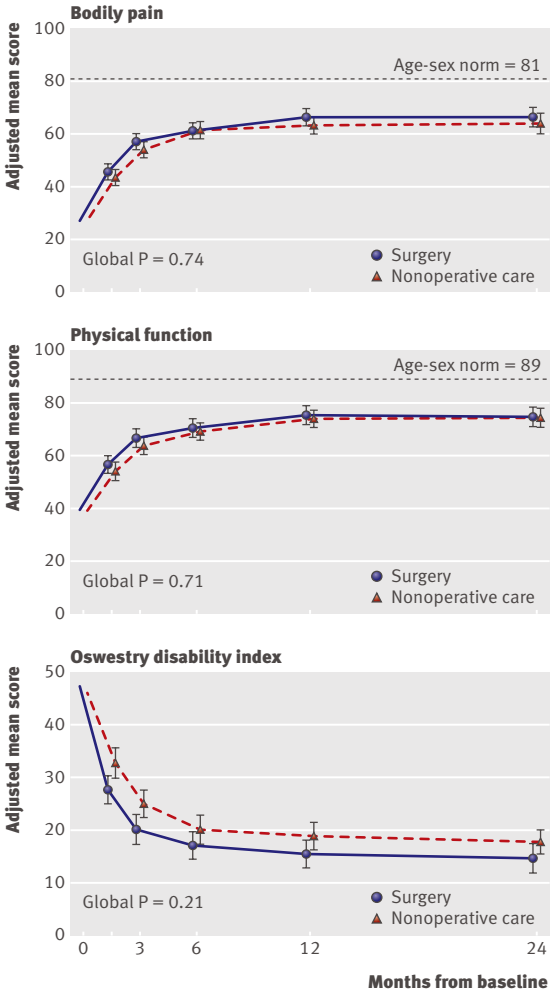


Fig 7 | Means scores over time for SF-36 bodily pain and physical function scales and Oswestry disability index from a study of surgical versus non-operative treatment for lumbar disc herniation (redrawn from Weinstein et al[22])

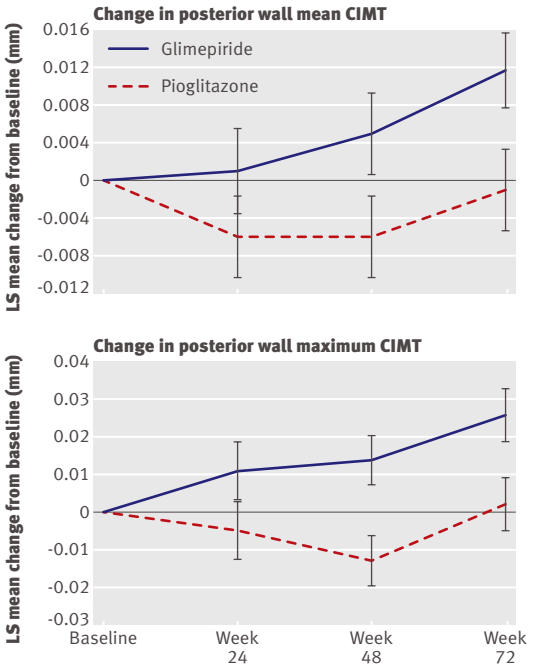


Fig 8 | **Changes in carotid intima-media thickness (CIMT) over time (redrawn from Mazzone et al[24]). Values are least square means using last observation carried forward and error bars are standard errors**

last observation carried forward is less desirable than an appropriate repeated measures model. Some trial reports may plot medians or percentages in certain categories rather than means over time, especially if the data are skew or categorical in nature.

Assessing visual evidence for a treatment difference

When a figure compares two treatments it is useful for readers to infer how the (lack of) overlap between standard errors or between confidence intervals indicates the strength of evidence for a treatment difference. The limits of a 95% confidence interval are about twice the standard error, slightly more if samples are small for a quantitative outcome. The following rough guide works well when two treatment groups have similar standard errors, which is often the case. Any overlap between the standard error bars means the difference is not significant. If there is a gap between the standard error bars that exceeds one standard error then the difference is significant, at $P<0.035$ in fact. Thus, a smaller gap may fall short of conventional significance. No overlap between 95% confidence intervals indicates strong evidence of a difference ($P<0.006$). So, a slight overlap between two 95% confidence intervals may still be significant.

It is important to note whether error bars are standard errors or confidence intervals, and to remember that displays of individual variability (such as standard deviations or interquartile ranges) do not help directly in detecting treatment differences.

Although figures are an important aid to interpreting results, the visual impression of a clinically relevant treatment effect needs clarifying by formal statistical evidence. We hope that the above advice will help readers to spot any deficiencies in figures and make their own wise interpretation of trial results.

What makes a good figure?

In our survey, figures were rarely explicitly misleading but some improvements could do much to enhance clarity. To this end all figures should:

- Emphasise clarity and be oriented to their primary goal
- Be independently interpretable
- Display measures of uncertainty when any estimates are plotted.

SUMMARY POINTS

Clinical trials contain four main types of figure: flow diagrams, Kaplan-Meier plots, forest plots, and repeated measures plots

Many published figures have deficiencies in presentation or content

Examples highlight good practice and pitfalls to avoid when interpreting figures

All figures are reproduced with permission from the original journals.

Contributors and sources: SJP originally formulated the project's broad intent. All authors then designed and carried out the survey. SJP wrote the article, TGT made major revisions following *BMJ* feedback and LMW helped to improve each draft. SJP acts as guarantor.

Competing interests: None declared.

Provenance and peer review: Commissioned; externally peer reviewed.

1 Tufte ER. *The visual display of quantitative information*. Cheshire, CT: Graphics Press, 1983.
2 Cleveland WS. *The elements of graphing data*. Summit, NJ: Hobart Press, 1994.
3 Gelman A, Pasarica C, Dodhia R. Let's practice what we preach: turning tables into graphs. *Am Statistician* 2002;56:121-30.
4 Schriger DL, Cooper RJ. Achieving graphical excellence: suggestions and methods for creating high-quality visual displays of experimental data. *Ann Emerg Med* 2001;37:75-87.
5 Puhan MA, Riet G, Eichler K, Steurer J, Bachmann LM. More medical journals should inform their contributors about three key principles of graph construction. *J Clin Epidemiol* 2006;59:1017-22.
6 Lang T, Secic M. Visual displays of data and statistics. In: *How to report statistics in medicine*. 2nd ed. Philadelphia: American College of Physicians, 2006:349-92.
7 Moher D, Schulz KF, Altman DG, for the CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657-62.
8 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663-94.
9 Schriger DL, Sinha R, Schroter S, Liu PY, Altman DG. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British Medical Journal. *Ann Emerg Med* 2006;48:750-6.
10 Pocock SJ, Travison TG, Wruck LM. Figures in clinical trial reports: current practice and scope for improvement. *Trials* 2007;8:36.
11 Edwards RD, Moss JG, Lumsden MA, Wu O, Murray LS, Twaddle S, et al. Uterine-artery embolization versus surgery for symptomatic uterine fibroids. *N Engl J Med* 2007;356:360-70.
12 Hay EM, Foster NE, Thomas E, Peat G, Phelan M, Yates HE, et al. Effectiveness of community physiotherapy and enhanced pharmacy review for knee pain in people aged over 55 presenting to primary care: pragmatic randomised trial. *BMJ* 2006;333:995.
13 Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet* 2002;359:1686-9.
14 Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA* 2007;298:1010-22.
15 Hochman JS, Lamas GA, Buller CE, Dzavik V, Reynolds HR, Abramsky SJ, et al. Coronary intervention for persistent occlusion after myocardial infarction. *N Engl J Med* 2006;355:2395-407.
16 Thiam S, LeFevre AM, Hane F, Ndiaye A, Ba F, Fielding KL, et al. Effectiveness of a strategy to improve adherence to tuberculosis treatment in a resource-poor setting: a cluster randomized controlled trial. *JAMA* 2007;297:380-6.
17 Cuzick J. Forest plots and the interpretation of subgroups. *Lancet* 2005;365:1308.
18 SAFE Study Investigators. Effect of baseline serum albumin concentration on outcome of resuscitation with albumin or saline in patients in intensive care units: analysis of data from the saline versus albumin fluid evaluation (SAFE) study. *BMJ* 2006;333:1044.
19 Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
20 Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analysis in clinical trials. *New Engl J Med* 2007;357:2189-94.
21 Smith I, Procter M, Geller RD, Guillaume S, Feyereislova A, Dowsett M, et al. Two year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet* 2007:369:29-36.
22 Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the spine patient outcomes research trial (SPORT): a randomized trial. *JAMA* 2006;296:2441-50.
23 Fitzmaurice GM, Laird NM, Ware JH. Adjustment for baseline response. In: *Applied longitudinal analysis*. New York: Wiley, 2004:122-32.
24 Mazzone T, Meyer PM, Feinstein SB, Davidson MH, Kondos GT, D'Agostino RB Sr, et al. Effect of pioglitazone compared with glimepiride on carotid intima-media thickness in type 2 diabetes: a randomized trial. *JAMA* 2006;296:2572-81.