Item 7a. How sample size was determined

Examples—“To detect a reduction in PHS (postoperative hospital stay) of 3 days (SD 5 days), which is in agreement with the study of Lobo et al17 with a two-sided 5% significance level and a power of 80%, a sample size of 50 patients per group was necessary, given an anticipated dropout rate of 10%. To recruit this number of patients a 12-month inclusion period was anticipated.”

“Based on an expected incidence of the primary composite endpoint of 11% at 2.25 years in the placebo group, we calculated that we would need 950 primary endpoint events and a sample size of 9650 patients to give 90% power to detect a significant difference between ivabradine and pla‑ cebo, corresponding to a 19% reduction of relative risk (with a two-sided type 1 error of 5%). We initially designed an event-driven trial, and planned to stop when 950 primary endpoint events had occurred. However, the incidence of the primary endpoint was higher than predicted, perhaps because of baseline characteristics of the recruited patients, who had higher risk than expected (e.g., lower proportion of NYHA class I and higher rates of diabetes and hypertension). We calculated that when 950 primary endpoint events had occurred, the most recently included patients would only have been treated for about 3 months. Therefore, in January 2007, the executive committee decided to change the study from being event-driven to time-driven, and to continue the study until the patients who were randomised last had been followed up for 12 months. This change did not alter the planned study duration of 3 years.”

Explanation—

For scientific and ethical reasons, the sample size for a trial needs to be planned carefully, with a balance between medical and statistical considerations. Ideally, a study should be large enough to have a high probability (power) of detecting as statistically significant a clinically important difference of a given size if such a difference exists. The size of effect deemed important is inversely related to the sample size necessary to detect it; that is, large samples are necessary to detect small differences. Elements of the sample size calculation are (1) the estimated outcomes in each group (which implies the clinically important target difference between the intervention groups); (2) the α (type I) error level; (3) the statistical power (or the β (type II) error level); and (4), for continuous outcomes, the standard deviation of the measurements. The interplay of these elements and their reporting will differ for cluster trials40 and non-inferiority and equivalence trials.39 Authors should indicate how the sample size was deter‑ mined. If a formal power calculation was used, the authors should identify the primary outcome on which the calculation was based (see item 6a), all the quantities used in the calculation, and the resulting target sample size per study group. It is preferable to quote the expected result in the control group and the difference between the groups one would not like to overlook. Alternatively, authors could present the percentage with the event or mean for each group used in their calculations. Details should be given of any allowance made for attrition or non-compliance during the study. Some methodologists have written that so called under‑ powered trials may be acceptable because they could ultimately be combined in a systematic review and metaanalysis, and because some information is better than no information. Of note, important caveats apply—such as the trial should be unbiased, reported properly, and published irrespective of the results, thereby becoming avail‑ able for meta-analysis. On the other hand, many medical researchers worry that underpowered trials with indeterminate results will remain unpublished and insist that all trials should individually have “sufficient power.” This debate will continue, and members of the CONSORT Group have varying views. Critically however, the debate and those views are immaterial to reporting a trial. Whatever the power of a trial, authors need to properly report their intended size with all their methods and assumptions. That transparently reveals the power of the trial to readers and gives them a measure by which to assess whether the trial attained its planned size. In some trials, interim analyses are used to help decide whether to stop early or to continue recruiting sometimes beyond the planned trial end (see item 7b). If the actual sample size differed from the originally intended sample size for some other reason (for example, because of poor recruitment or revision of the target sample size), the explanation should be given. Reports of studies with small samples frequently include the erroneous conclusion that the intervention groups do not differ, when in fact too few patients were studied to make such a claim. Reviews of published trials have consistently found that a high proportion of trials have low power to detect clinically meaningful treatment effects. In reality, small but clinically meaningful true differences are much more likely than large differences to exist, but large trials are required to detect them. In general, the reported sample sizes in trials seem small. The median sample size was 54 patients in 196 trials in arthritis, 46 patients in 73 trials in dermatology, and 65 patients in 2000 trials in schizophrenia. These small sample sizes are consistent with those of a study of 519 trials indexed in PubMed in December 2000 and a similar cohort of trials (n=616) indexed in PubMed in 2006, where the median number of patients recruited for parallel group trials was 80 across both years. Moreover, many reviews have found that few authors report how they determined the sample size. There is little merit in a post hoc calculation of statistical power using the results of a trial; the power is then appropriately indicated by confidence intervals (see item 17).