21.3: Statistical analysis plan

A common mistake in the planning of a trial is to delay consideration of the analyses until the data become available. It is essential that the main analyses that will be undertaken are planned at the design stage, as this provides several major benefits. First, it encourages a clearer understanding of the basic questions to be answered and thus assists with the formulation of clear and specific objectives. For example, in a vaccine trial, a simple comparison of the numbers of cases of the disease occurring over a 5-year period in the vaccinated and unvaccinated groups may answer the question of the magnitude of any protective effect. A comparison of the incidence rates of disease in vaccinated and unvaccinated individuals in the first, second, third, fourth, and fifth years after vaccination can be used to answer a rather different question, namely, whether the protective effect is constant over the 5-year period.

A second benefit of considering the analyses at the design stage is that it necessitates specification of what data need to be recorded. The investigator can check that arrangements have been made to measure and record all variables that will be needed in the analyses. Also, and perhaps as importantly, it may become clear that some variables will not be needed, and these can then be omitted from the study.

The process of planning the analyses may identify also the importance of subgroup analyses. In a vaccine trial, for example, it may reveal a need to assess the efficacy of the vaccine in children vaccinated at different ages. This may have major implications for the choice of sample size, as the need for age-specific estimates of efficacy requires a much larger sample in each age group than would be needed if only an overall estimate of efficacy was wanted.

Finally, advanced planning of the analyses is desirable to ensure that adequate arrangements have been made for data handling, the necessary computer software is available, and sufficient time for data cleaning and analysis has been allowed for in the study schedule.

Prior to any formal statistical analyses of the kinds discussed from Section 4 onwards, it is essential to perform simple tabulations of data and to construct simple diagrams to summarize the information that has been collected. Simple
statistical package computer programs, such as Epi-Info (<www.cdc.gov/epiinfo>) or STATA (<http://www.stata.com>), greatly facilitate doing this. The investigator should use these simple approaches to gain a good understanding of the data collected, before embarking on more complex analyses. These simple analysis methods are not described further in this manual, but they are discussed in most good textbooks on medical statistics (for example, Armitage and Berry, 1987; Kirkwood and Sterne, 2003).

If the results of a trial are to be used for submission to an appropriate authority to grant a licence for a new drug or vaccine, the licensing authorities will require that a statistical analysis plan (SAP) is developed as a separate document, to be completed after finalizing the protocol and before the code is broken for who is in the intervention and control groups (if it is a blinded trial). The SAP should contain a technical and detailed description of the principal analyses to be conducted on the trial data, which has more detail than would typically be included in the trial protocol. The plan should include detailed procedures for conducting the statistical analysis of the primary and secondary outcome variables and of other relevant data. Often, the licensing authority will require a copy of the SAP for them to examine and approve in advance of a trial being analysed.

It is good practice to prepare a SAP for any trial, even if the results are not to be used for product licensing. In addition to any necessary review by licensing authorities, the SAP should be reviewed and approved by the trial steering committee and also often by the trial data safety and monitoring committee (DSMC). A formal record should be kept of when the statistical analysis plan was finalized, as well as when the final data set was ‘frozen’ and when the trial was unblinded.

It is common to develop the computer programs for conducting the SAP in advance of breaking the treatment code. To check that these are working properly, some analysts assign study participants at random to intervention or control groups (irrespective of which group they were actually in) and run the programs on these ‘test’ data. In this way, they are able to check that the final tables are in an appropriate format to be interpreted, once the code is broken. Conducting such a ‘dummy run’ analysis generally greatly speeds the analysis and interpretation of the trial, once the data are finalized.

Often, when an analysis is conducted, further analyses will be appropriate and prompted by an initial examination of the study results, rather than being pre-planned in the SAP. Such analyses are often called ‘exploratory’. They were not specifically planned in advance but were prompted by examination of the trial findings. Such exploratory analyses are sometimes informative and may suggest new hypotheses, but it is important to distinguish them from the analyses that were included in the SAP, as they were suggested by the data, rather than being planned in advance of the code being broken. It is generally wise to interpret the results of such exploratory analyses with caution.