20.1: Introduction to data management

All intervention trials involve the collection and management of data, often in large quantities. In order to get the most out of study data, it is important to have worked through plans for the collection, management, and use of the data early in the planning stages of a trial. Previous editions of the Toolbox discussed the role and choice of computers in the management of trial data, but now they are so ubiquitous that there will be few trials in which they are not central to data handling and analysis. Indeed, developments in computing have changed the way that trials are conducted, from the way that data are collected through to the way data are used and disseminated. However, in the processing of data, it is important to remember the ‘GIGO’ principle ‘garbage in, garbage out’! The data used in final data tables are only as good as the data that go into their construction. Thus, while developments in computer hardware and software have made the processing and analysis of data much quicker, it is still necessary to pay careful attention to the way in which the original data are collected and recorded in the field and transferred from one program to another during the data management process. Every instrument used in the study, including questionnaires, laboratory methods, and data management programs, must be properly validated and tested and have good quality control (QC) procedures in place throughout the trial. Great attention to detail is necessary in every step the data take, in the design of data forms, in the recording of data in the field, in transferring the data from paper to the computer (if data are not collected digitally), in the transfer from one software package to another, and in how they are manipulated and managed in computer packages and programs. These data processing aspects are the focus of this chapter. The chapter focuses exclusively on quantitative data.

Section 2 covers some of the data-related issues that should be resolved before the study starts, and Section 3 concerns the planning that should be done for the data flow within the study. Sections 4 to 7 deal with various specific issues related to data flow and data management. This chapter can only give a basic introduction to key issues related to data management. More detailed explanations are available in various books and other resources. The general principles of data management are covered in books by Hernandez (2013), Powell (2006), McFadden (2007), Murrell (2009) (available free via <www.stat.auckland.ac.nz/~paul/itDT/>), Prokscha (2012), and Pryor (2012). Other free online resources include...
resources are provided for specific data management software, such as Epi-Info (<http://www.cdc.gov/epiinfo>) and EpiData (<http://www.epidata.dk>), or Microsoft AccessTM (<http://office.microsoft.com/en-us/access/>) such as for Access 2007 (<office.microsoft.com/en-us/access/HA012242471033.aspx>), and there is a useful web-based discussion group for data managers within the Global Health Trials website (<http://globalhealthtrials.tghn.org/community/groups/group/data-management-statistics/topics/290>).