20.4: Data collection systems

In this section, we review some of the ways in which data can be collected from the participants and put into an electronic database.

4.1 Questionnaires

Paper-based questionnaires are often used to capture responses from study subjects, especially in small studies. These will need to be printed, taken to the study site, collated, batched for data entry, stored, and preserved for future reference. The design of questionnaires is discussed in Chapter 14.

4.2 Electronic data capture

Electronic data capture through the use of field computers, PDAs, UMPCs, or mobile phones is increasingly used. Using electronic data capture makes the data available immediately and removes the need for separate data entry, but it increases the need for data quality checks at the time of data collection. Electronic data capture devices need to be programmed to ensure that checks on the data quality are performed at the time of collection, as it is difficult to verify the data afterwards. With electronic data capture, it is easier for additional modules to be administered to a sub-sample of participants. These additional modules can be triggered by specific questions, for example, loops to ask about all the children in the household or about all the medicines taken at the last illness.

Using electronic data capture properly can enable data to be collected quickly and allows for numerous checks of data quality to be built in at the time the data are collected. Open source software exist for many applications, such as openXdata (<http://www.openxdata.org>), OpenEHR (<http://www.openehr.org>), and ODK (<http://opendatakit.org>), with the advantage that source code is available for modifying and adapting them.
Collecting data using mobile phone applications is becoming increasingly common. Mobile phones are relatively cheap, and telecommunications network coverage in most countries makes them available to large sections of the population. Information can be collected remotely, wherever the study subject might be, and the person does not have to be questioned face to face by an interviewer. Mobile phones can also be used to collect repeated data from individuals who may be difficult to locate or who may be in remote locations. Computer programs, such as FrontlineSMS or EpiSurveyor, allow data to be collected through simple text message or through interactive voice response or self-administered questionnaires. In all these cases, the data are stored directly in a central database, following transmission across the telephone network, and are available for processing almost immediately, following collection.

4.3 Laboratory data

Data from laboratory tests are important in many research studies, and it is important to design the stickers, labels, and linking mechanisms, so that samples collected in the field can be linked to the results of the laboratory tests and the other data collected on the same individual. Many laboratories use laboratory data management systems (LDMS), such as LIMS, which automatically download laboratory results into a computer database (see Chapter 17). Alternatively, the results can be entered onto paper or electronic forms, which are later merged into the database.

It is better to use a unique specimen identification number, rather than the individual’s study identification number. This is because a single individual may have several specimens of the same type taken during a trial. As a check, the questionnaire number should be written (or a sticky label can be used) on the laboratory form, and a copy of the specimen identification sticker placed on the questionnaire, as well as on the specimen itself. If both the individual’s study identification number and the specific specimen identification number are used on both forms, there can be assurance that the samples are correctly matched to the questionnaires when the analysis is done. Bar codes can be used for these laboratory numbers to enable the code to be read automatically by the laboratory equipment (also see Chapter 17).

LDMS must be programmed and managed carefully. Often, several studies use the same laboratory for many different tests. The LDMS must allow a study team to access all the data for their study, but they must not be able to access data from other studies. This requires common protocols and database programs, and good SOPs to ensure that data access is controlled and monitored.

4.4 Clinic data

Data from hospitals and clinics are sometimes used in trials. Patient-level data may be collected by clinicians when they assess, diagnose, and treat patients who are participating in the trial. The clinical data can be collected on a separate dedicated form, from which data are entered into the database later. Alternatively, there may be a trial research assistant in the clinic who enters the data into the computer from clinical records, or an electronic data collection tool may be introduced for use by the clinician, which removes the need for paper forms. With suitable choices for database programming and hardware, such systems can be relatively cheap and cost-effective.

There are several software options for the collection and management of health records from clinics and other health facilities (such as openXdata, openEHR, open-MRS). These support data entry at the time the patient is seen by the clinician.
4.5 Longitudinal data collection

Longitudinal data require a system to link individuals within the database with each of the occasions when they are followed up. To do this, personal information, such as the person’s names, address, and/or an identity number, needs to be stored in the database and used for subsequent survey visits to make a positive identification of the study subject. In order to make the identification more certain, photographs of the study subjects or fingerprints might be collected. These methods are cost-effective for even small studies, using mobile technology such as PDAs, cameras, and mobile phones.

The first time any individual is seen, sufficient personal information must be collected at the time that they are assigned a unique study identification number, so that unambiguous identification can be made on the second and subsequent visits.

Such personal identifiers must be kept secure and confidential, especially if these can be linked with health information or other sensitive data. However, appropriate information for identifying individuals must be made available to the fieldworkers at follow-up visits, through printed lists or through access to the electronic database, using PDAs, UMPCs, or other mobile computing devices. Links between the study numbers of individuals who belong to the same family or household can be easily stored in relational databases.

4.6 Quality control

In all trials, there is a need to ensure the quality of the data collected. To do this, it is necessary to be able to answer, and show evidence for, the following questions. Are the data a true reflection of the response from the study subjects? Has anyone changed the data and, if so, how? Is there effective QC over the data collection and data management? Are the data correctly matched and linked to the right respondents? (See also Chapter 14.) It is important to build quality checks and audits into the data collection and their subsequent management, in order to have the evidence to answer these questions. These checks fall into four main areas: design, training, supervision, and checking. Data collection should build in design features that allow checks and simplify coding and responses. Training should include a thorough examination of the instructions that all data collectors should know and follow. It should also explain and go over the ways that the data are checked at all levels, so that everyone knows that the process has checks and balances and that mistakes will be found and corrected. Supervision is important and should be supportive and non-threatening, with the objective of building quality and encouraging self-assessment and improvement. Regular tallies should be kept of the number of questionnaires completed, the number of refusals, and the number of errors or mistakes discovered. At the beginning of data collection, daily tallies of these indicators may be needed, but even weekly or monthly tallies may ensure that difficulties with the data collection are picked up early, and re-training given to those who need it.

Audit trails are used to keep track of any changes in the data. While every effort should be made to collect the correct data at the time of the interview or measurement, there will always be times when data need to be changed. Before the advent of computers, data managers used to keep logs of their work in ledger books, recording all the changes made to the database. Now any changes that are made should be documented in the database, which will include a record of the old values and a record of the reason for the change. Computers should never be programmed to make changes automatically. Rather they should be programmed to highlight probable errors, and a data manager can make any necessary changes and record the reasons for each of the changes.
4.7 Future trends

The traditional ways of collecting data through paper-based questionnaires will continue to be needed for some studies, but there are increasingly diverse other methods available. The use of mobile phones for collecting data has grown substantially in recent years. They have the advantage of enabling data to be collected frequently, and at any location or time, but currently are limited in the amount of data that can be collected at any one time.

Computer-assisted self-interviewing is a growth area. The advantages are that questions are standardized and confidential, and many people can be interviewed at the same time. Translations of questions can be made into different languages. The questions can be delivered in many ways, as an audio system for those who cannot read or through pictures and visual choices available through touch-screen technologies.

Online databases have become much more accessible and allow direct data collection into a master database located in the study centre or elsewhere. Mobile phone networks allow instantaneous transmission of data from the field to the data centre where it can be checked against the master database. Based on the data sent to the online database, fieldworkers collecting the data can be given instructions about the data to be collected and new study subjects to interview. These systems are increasingly used by large multicentre studies but will become more applicable to smaller studies where the online database can be linked to other resources, in order to improve the study design or data collection.