20.2: Before starting to collect data

All trials need appropriate resources to collect data and information, to check the consistency and quality of the data, and to organize the data into a suitable form for analysis. It is important that all the steps of the trial and the associated data flow are planned before starting the trial, and the resources needed at each step are defined. This section describes the different hardware, software, personnel, and systems needed to process data in a trial. When considering the trial budget, resources must be allocated for all of these aspects, and often components have to be capable of multitasking, for example, computers that can be used for both data entry and administrative functions, and software that can manage different data formats.

There are four components to the description of the data processing for a trial:

1. hardware, i.e. any physical entity used for data processing. This may include computers, printers, and electronic hardware, but also includes paper, pens, and other equipment used to collect, transfer, and archive the data
2. software, i.e. the programs needed to make the hardware manipulate and process the data for the study
3. personnel that are needed for the data processing
4. the systems and organization that must be in place to bring all of the different components together.

2.1 Hardware

The commonest hardware used in a small trial is still paper questionnaires and forms. Much of what is done is recorded on paper, and, at all stages, paper copies are kept as the definitive record. The advantage of using paper is that it is a physical entity, which preserves the data content. The disadvantage is that it is difficult to process and analyse, particularly in large quantities. Data collected or stored electronically are much easier to manipulate and use in a variety of different ways.
If paper systems are used to collect data, it is important to include, in the planning, provision of all the necessary ancillaries for the paper collection such as pens, clipboards, and storage boxes. Management of the paper is also an issue that needs to be thought through to the end of the trial and beyond, with proper filing systems and archives for data storage. Paper systems need to be integrated with the computer hardware and software used in the study, first to do the printing of the questionnaires and other forms, and second to take the data from the forms and input them into a computer package for electronic checking and analysis.

The use of computers to collect, process, and analyse data is ubiquitous nowadays. There are so many different computers, and they are continually getting better and faster that it is impossible to give very specific guidance on which would be best for particular studies. Much depends on the way the data management for the study has been planned and what software the analyst is already familiar with. One way to divide up the many computer hardware options is through the distinction between desktops, laptops, mobile devices, and servers. Desktops are useful for data entry and when there are many people wanting to share a computer for short periods of time, for example, field supervisors who need to input a report at the end of the day. Desktops are also needed for some of the administrative functions, but, in general, it is good to keep the research data physically separate from the administrative computers that the project needs.

Laptops provide comparable computing power to desktops and can be used to collect and manage data in the field, even where mains electricity is not widely available. Smaller devices, such as PDAs or ultra-mobile personal computers (UMPCs) or even ‘smart’ mobile phones, are easier to use and transport in the field than laptops. With laptops and smaller mobile computing devices, two issues need to be considered; first, the smaller the device, the easier it is to lose or be stolen, and second, these devices have batteries that need recharging and periodically replacing. When purchasing laptops and smaller devices, buying a security cable for each machine, where appropriate, is often a good investment. It is also important to make sure that the person responsible for the computer uses the security cable and the procedures are well known to all, as it only takes a minute to lose large amounts of data if a computer is stolen. If in continuous use, recharging laptops, PDAs, UMPCs, or mobile phones can be a time-consuming task. Long-life batteries can be used to extend the time the machines can be used between charging, and, if mains power is available, recharging can be done at meal breaks and overnight. Otherwise, inverters can be used to charge from car batteries or from solar panels.

PDAs, tablet computers, mobile phones, and other devices can be programmed to accept electronic questionnaires and can also be purchased with GPS software, cameras, bar code readers, and automatic Internet capabilities, with the only drawback being the cost of the extra functions. In general, it is important to specify what is needed for the trial and to avoid expenditure on functions that are not needed.

All but the smallest trials will benefit from having a server, in order to store the data and to manage resources. A server can be a special computer with a large amount of data storage capacity or a standard desktop configured to organize data storage and administrative procedures. However, servers do need to be looked after carefully, with control of the temperature, dust, and humidity in the server room. If the trial operates out of an established institute, it is likely that it will be possible to use the institution’s server and network, perhaps through creating a virtual server for the use of the trial. The networking of the server can be through physical cables or could be set up as a simple local area network (LAN) using a wireless router, but note that, while laptops usually have built-in wireless capability, this is often not the case for desktops and PDAs. A good server and network can simplify many operations, such as access to the Internet and sharing of data, and should be high on the list of priorities for all but the smallest study.
Ancillary equipment (‘peripherals’) is also needed. This may include printers, scanners, photocopiers, cameras, bar code readers, and backup devices. These can be installed and connected to one computer, but a simple network will make it easier for different members of the research team to access the different peripherals. The wider access to the Internet needs to be planned as well. If the Internet service is poor, it may be necessary to have more than one way of accessing the Internet, perhaps through fixed lines or through mobile phone networks.

### 2.2 Software

In this section, we will not consider general software, such as word processing or anti-virus, which are typically available to all computers, but concentrate on the specific options available for data processing. Data processing software comprises specialist packages which facilitate the collection, management, and organization of trial data. They can be used to prepare data for analysis by specialist data analysis packages. We consider three broad categories of software—freeware (free software packages), proprietary software (which must be bought), and open source software—and give some examples of the different packages available, but the choice is wide. The most important consideration is to plan out how the data processing for the trial will be done and to use the appropriate package for each step in the data flow. It should be simple to transfer data from one package to another and is wasteful of time and resources to do any data operation in a package that is not designed for that purpose. In many ways, the selection of the software is more important than the hardware, and good selection can save a lot of time.

For the sort of data that are collected in epidemiological studies and trials, Epi-Info (<www.cdc.gov/epiinfo>) is a very useful freeware package which can be used for many types of study. A similar freeware package Epi-Data (<http://www.EpiData.dk>) provides data management, analysis, and transfer capabilities. These packages are easy to learn and use and are ideal for small studies.

For larger studies, it is usually better to use a proprietary software package such as Microsoft Access® or MS-SQL®. These are easy-to-use software, with good learning materials to help in developing and using the database. These software packages can be used to clean and manage data, and it is easy to transfer data from them to analysis packages. Free, but limited, versions of these software packages are available for those on limited budgets.

Open source software packages are also usually free to the user, and it is possible to access the source code and develop applications that are tailored to specific studies. A challenge, however, is learning how to manipulate the source code and make the software function appropriately for a specific study. Examples include RedCap which is aimed at investigators who do not have access to much computing support but who wish to quickly set up and manage clinical studies, including longitudinal ones, while OpenClinica targets researchers conducting clinical trials that must meet the regulatory requirements of the US Food and Drug Administration. Open Data Kit (ODK) is a suite of open source applications that allow the creation of questionnaires for data collection on Android-enabled mobile devices and facilitate online data management. Force.com is a powerful data management platform, for which a limited number of free licences are provided to non-profit organizations and higher education institutions. All of these packages are free to use and are highly customizable, and all but Force.com can be configured without highly specialized computer programming skills. All four systems are supported by knowledgeable end-user-driven online communities.
2.3 Personnel

The personnel needed for data management will depend on the size of the trial and the computer and software systems being used. For a small study, using paper forms and a simple software package, such as Epi-Info, for data processing, one part-time data manager and one data entry clerk may be sufficient. For larger studies using paper forms, a team of data entry clerks and several data managers might be needed. Although the requirements for data entry clerks can be greatly reduced or eliminated for studies using electronic data capture, skilled data managers and expert programmers may be needed to program some of the collection devices to validate the systems and to design the database.

Successful data management requires a variety of different skills at different times during the study. At the start of the study, except for simple software packages such as Epi-Info, someone with technical skills will be needed to set up the database and write the data check programs. If the study personnel are not skilled in database programming, it may be better to hire a consultant to do it. When the study is underway, staff will be needed to enter data (unless all data are captured electronically), manage data checking, and clean the data on a daily basis. By the end of the study, data files must be prepared for the statistical analysis and report writing, and again buying in the expertise may be appropriate if there are no staff in the team with the necessary skills. Depending on the size of the study, some of these roles may be combined in a single individual, whereas, in larger research groups, individuals may be specialists who work full-time in one area of data management such as database development or writing data checks.

Staff must be recruited before the trial starts to be able to both develop, and be trained in the use of, the data processing systems. It is important to allocate sufficient time for training. Even staff who have previous experience of data management on other studies will need to be trained to use the system being used in the trial and become familiar with the study protocol. The most important attributes for data entry staff are conscientiousness, reliability, and attention to detail. Existing computing skills may be less important, as staff can be trained to use a computer and to enter data. Sometimes, staff who were originally employed to collect data in the field can be trained to be good data entry clerks. This has the advantage that they will be familiar with the kind of data being collected and the forms in use. They will also be aware of the problems that may arise in the collection of data in the field. However, data entry clerks and their supervisors are gradually being reduced in number in many research groups, as they move from collecting data on paper to electronic data capture.

A supervisor is likely to be necessary for every four to six data entry clerks to control the quality of their work, to ensure a proper and equitable distribution and flow of work, and to ensure that all data and forms are correctly processed and stored. The supervisor may be able to do some of the initial data checking and cleaning and take some of the data management tasks from the study data manager. A good way to identify persons who might be trained as supervisors may be to select them from among the data entry clerks, based on their performance and aptitude for this work, although the ability to type data quickly and reliably does not necessarily provide a good indication that an individual will make a good supervisor.

Pilot studies may be necessary to determine how much data can be processed by a data clerk in a day, to know how many such individuals to include in the trial budget. This should be part of the pilot testing, which is covered in more detail in Chapter 13. As the work is repetitive, but requires considerable care, it is advisable to plan that a clerk should not be entering data for more than 5 or 6 hours a day. Data entry may be interspersed with filing tasks to maintain variety in the work.
If a trial is large, substantial numbers of forms may accumulate quickly, and the design of an appropriate filing and tracking system, such that individual forms can be retrieved, if needed, is important. The employment of filing clerks may also be necessary in large trials. Data entry and filing are tasks that need to be done in the same way, day after day. So it is important to devise ways of maintaining staff morale, so as to ensure high-quality work. For larger trials conducted over several years, working out career development structures within the project may be important (for example, the progression from filing clerk or fieldworker to data entry clerk to supervisor). Also, training in new techniques and the use of computer packages may be appropriate. Individuals must be aware that their work is considered important and that its quality is monitored, so that bad work is detected and will need to be corrected, while good work is noticed and rewarded appropriately.

The data management staff must be made to feel that they are an integral part of the project. Appropriate measures should be installed to allow field and data management staff to liaise with each other, so that they consider themselves part of the same team. See Chapter 16 for more details of field operations. Field staff must understand the problems that errors in data collection cause in the processing and analysis of data, and data management staff must appreciate the obstacles to high-quality data collection in the field. Visits by data staff to the field can do much to aid such mutual understanding, as can field staff spending short periods working or observing in the data office.

2.4 Data oversight

No matter how good data systems are, there is always benefit in getting someone outside of the study to look at them to see if they can be improved. The best time for this is before starting to collect real data, in time for the systems to be changed, if necessary. For small studies, this may be a matter of getting a colleague to check the data systems. In larger studies, outside advisors might be hired to look at the data system.

In most clinical trials, the requirements of good clinical practice (GCP) are such that the trial data must be collected and processed, in compliance with the ICH–GCP guidelines (see Chapter 16, Section 7.1). The practical implications of these requirements are that the data management process must be documented, and the computer systems used to collect, store, and process the data must be validated. The regulations governing the management of data from clinical trials can be broadly classified into: (1) clinical data-related; (2) technology-related; and (3) privacy-related. The guiding principle behind all of these regulations is the need to be confident that the data were collected as defined in the study protocol, are from real participants, and can be independently verified. Small studies may not be required to implement GCP, but, for all studies, there should be awareness that good practice and procedures should be in place, ensuring that data systems are checked for errors or oversights.

Compliance with GCP requires that all phases of the data management processes are controlled by standard operating procedures (SOPs). Data management staff must be trained in each process, and training must be documented. ICH–GCP does not require double data entry but requires that processes are in place to ensure that the data in the database accurately reflect what was recorded in the field on questionnaires or through other means.

The computer system used to store and manage the data will need to be validated, which requires a validation plan, user specification, testing, and change control. In the simplest form, an SOP that describes the steps necessary to build, test, and release a database can serve as the validation plan. The database and data entry screens will need to be tested to ensure that they function correctly, and the testing and its results should be documented.
A further requirement of GCP is that all changes that are made to the data in the study database are documented and that the original data are not deleted. This requirement is generally interpreted to mean an electronic audit trail must be created, in which the software system automatically records any changes that are made to the database, including when they were made and who made them. However, there are differences in how the term 'audit trail' is interpreted and implemented and at what stage the audit trail is ‘turned on’. Some audit trails may record changes after first entry into the database, others after second entry when data have been verified, and others not until after initial data cleaning is done. Building a database with an electronic audit trail requires specialist skill and knowledge; however, software packages specifically designed for clinical trials, such as OpenClinica, have an audit trail as an inbuilt feature.

In a small trial that does not involve licensing of a pharmaceutical product, it may be possible to document data changes by other means to demonstrate compliance with GCP, for example, keeping a copy of the original database after second entry, a separate database containing all updates to the data, and a paper record of all changes that are made.

GCP also requires that a security system is maintained that prevents unauthorized access to the data. This would generally mean having a separate password to access the database and users having different levels of permitted access, depending on their role in the data management process. Randomization codes (see Chapter 11 for details) should always have restricted access, so that unauthorized staff cannot find out which treatment has been allocated to which participants.

GCP also requires that data are backed up adequately. Even in a study that is not being run to GCP, it is essential to develop a system for regular data backups. Failure to do so may result in the loss of data. Several types of media can be used for backup, including tapes, CDs, or external hard drives. Whatever is used, backup copies should be made regularly (at least weekly and possibly daily), once data entry has started. At least two backup copies of the database should always exist, and periodic ‘restores’ of the backed up data should be done to verify the data integrity. The copies should be updated regularly and frequently, although it is a good idea to keep some old versions as well, as errors are sometimes found in the more recent ones that make it necessary to restart data entry from a previous copy. Some of the copies should be stored in a geographically separate location in a dry and relatively dust-free environment (for example, in a sealed plastic bag). Complete records should be kept of the data that are stored on all backups, with one copy stored with the backup and at least one other copy stored in a separate place.