NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Smith PG, Morrow RH, Ross DA, editors. Field Trials of Health Interventions: A Toolbox. 3rd edition. Oxford (UK): OUP Oxford; 2015 Jun 1.

# Chapter 20 Data management

## 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 <https://www.stat.auckland.ac.nz/~paul/ItDT/>), Prokscha (2012), and Pryor (2012). Other free online resources are provided for specific data management software, such as Epi-Info (<http://wwwn.cdc.gov/epiinfo>) and EpiData (<http://www.epidata.dk>), or Microsoft Access™ (<http://office.microsoft.com/en-us/access/>) such as for Access 2007 (<http://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>).

## 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 (<http://wwwn.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 under way, 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.

### 2.5. Summary

In this section, emphasis has been given to the need to plan the data system and all its dependencies before starting the trial. This involves planning the hardware and how it will be used, the software and what it needs to be able to do, the personnel, and the data practices. If the data systems are planned and thought through at the beginning of the study, the study progress is less prone to error and easier to operate.

It is useful to make a process flow diagram of the data indicating the people who will handle the data at each point in the flow diagram. Taking the time to highlight the resources needed and the person responsible for each action can make the implementation easier when the study is under way.

## 3. Planning the data flow

There are many advantages to collecting and storing research data electronically. Electronic storage of data facilitates easy retrieval, simpler generation of study reports, easy exportation to statistical packages, and rapid data sharing. The benefits of electronic storage of data can only be fully realized if the database storing the data is well designed. A poorly designed database leads to poor performance, inefficient data queries, inaccurate and unreliable data, and redundant data that are duplicated in many places, making it difficult to check and clean. This section focuses on the key aspects of the processes in the data flow.

### 3.1. Database design

Database design is the process of organizing data in such a way that it can be stored and retrieved efficiently. It involves making decisions on how best to model a real-world information system, such as a paper-based data collection system, into a database. It is very unlikely that an analyst can correctly design a system without a full understanding of the key processes and activities involved in the study. This requires researchers spending time with the system developers to ensure that the system developed is what is required.

It is good practice to use a structured approach, referred to as a system development life cycle, when undertaking a database development project. The choice of the methodology is usually influenced by factors such as the complexity of the proposed database, the size of the database and the programming team, cost, time, and criticality of the project. What is important is to get an approach that meets the needs of the project. An overview of the key phases involved in database development is presented, rather than focusing on a specific methodology. These procedures should not been seen as checklists, but rather key processes that can be incorporated in any methodology chosen. The key procedures are:

1. project specification
2. requirements gathering
3. programming and testing
4. database implementation (‘going live’)
5. database maintenance and change management.

A database project should start by clearly defining what the database will be expected to do. The high-level requirement is defined, which is the mission statement that states the intended goal of the new database. It should not be more than a few lines long. Other critical factors to define at this initial phase are the scope, resources, timelines, hardware, software, and the database team. The scope is the boundary of the system and database, and it states what data and functionality will be included and what will be excluded. It is important that the scope is defined clearly at the start of the database development project, as poor definition leads to ambiguity and poorly defined database requirements. It is also important to choose the hardware and software early, as this may affect some of the design features of the database. The output of this initial phase is a project specification document that defines the objectives, timelines, deliverables, and milestones.

The objective of the next phase is to transform the high-level requirements into more detailed manageable tasks and functions that can be programmed into a software system. The requirements can be gathered by interviewing end-users of the proposed new database, examining the current database, if any, and also looking at existing forms such as questionnaires and reports. It is important to think of what extra functionality is required in the database. Will the data be shared? If so, which specific data will be shared? What are the security and compliance requirements? Have risks been assessed, and the database designed to mitigate the risks? The output of this process is a detailed requirements specification document, describing all the functional and non-functional requirements of the database. It is imperative that requirements are specified correctly and as comprehensively as possible at this phase; otherwise, it could lead to a system that does not meet its intended goal and may necessitate major changes during programming and testing or after the database has gone live.

The third phase involves creating a conceptual design (logical diagram) that shows the different tables that will be required to store the data identified in the first two phases. A good place to start from when generating the list of possible tables and their attributes (data columns) is to look at the current process, if any, used to collect data. There are two possible scenarios—an existing computer system is being converted or modernized or a new system is being built from scratch. In the former scenario, the tables and data entry forms of the existing computer system should be used as a starting point. If there was never a computerized system in place, begin with the existing paper-based data collection forms. If there are none, sketch out the forms, based on the requirements specification document, and discuss the sketches with the research team, and refine them further. Note that, while some of these data entry forms are sketches of what will eventually become data entry screens, others will properly remain in the realms of paper forms and will not necessarily map directly into data entry screens. If new requirements arise, while the conceptual design is being created, add them to the list of requirements that was created in the earlier phase. While sketching out the tables, also review the list of existing and new reports to establish a reasonably definitive list of the reports that the system must produce if it is to satisfy the needs of the users. The objective of analysing the reports is to ensure that the tables sketched out will have all the attributes that are needed to generate the reports. If there are missing attributes or tables, they should be added now. The completeness is important, but it can sometimes be difficult to know if all the reports that are being, or will be, produced or used have been identified. The database developer can proceed to create the physical database when the team considers that the requirements are sufficiently comprehensive.

After the database has been created, a programmer designs the data entry forms and links them to the tables in the database. There are various types of software that can be used to create the electronic data entry forms to capture the data. The choice of the tools and programming language will depend on the technical skills and preference of the team. When the programming phase is done, the database application should be tested. It is recommended that someone other than the programmer who developed it tests the application. Testing is an important phase, because it ensures that the system is validated and verified, a major requirement for GCP compliance. Users have to be trained on how to use the database, before deploying it for actual use. The database application should supplement user training by providing help features where users can access help through the application.

### 3.2. Data cleaning and integrity

Data cleaning should be an ongoing process, rather than something that is done at the end of the study. The process by which the data will be cleaned should be well thought out, planned, and documented at the beginning of the project, and certainly before any significant volume of data has been collected.

Double data entry is commonly used to minimize data entry errors. In this technique, two different people enter the same record independently, and the two entries are compared against each other. A validated data record is one where both entries are the same (see Section 5.1). It is important to remember, however, that no data entry system can avoid errors that were made by the interviewer using a paper-based questionnaire to record information in the field.

The database application can be programmed to flag inconsistencies in records, either during or after data entry. One approach is to categorize errors as being critical and non-critical. A critical error is one that is so important that systems are put in place to ensure that the data record cannot be saved into the database until the error has been fixed, for example, lack of the respondent's identity code or this being out of the valid range, as this code will be needed to link information in the database. The programmer would write these as checks embedded in the database application, sometimes called 'online checks'. The downside to having too many such checks embedded into the data entry screens is that the users cannot save the data until all the errors have been fixed, which can lead to back-logs. Decisions about what errors will be critical and non-critical should be made early enough, so that these are programmed in the system. Non-critical errors should not stop the user from saving the data record. They would instead be flagged up as data queries and reports for the data manager to follow up, rectify, and update the database.

Another approach is to incorporate data checks in a statistical program, for example, Stata, and run the checks against the data periodically. In the case of a paper-based collection system, periodic monitoring visits can be made to ensure that SOPs are being adhered to. Further checks can be done by taking a random sample of paper forms and comparing them against the corresponding electronic data records.

### 3.3. Programming issues

Computerized data collection systems are driven by computer programs written by system developers. The resources used to develop these systems can be made more effective if good programming practices are used. Computerized data systems should be documented to a sufficiently detailed level, so that any other system developer could quickly take over the maintenance or extension of such a system. A poorly documented computerized data system makes it very difficult to make changes to the existing system, and it will take a second person longer to figure out what needs to be changed. Even the programmer who wrote the initial program may forget specific technical details after a few months. Investing the effort to document programs, as they are developed, makes them easier to maintain, and changes can be made much more quickly.

Prototyping is an iterative technique used in computer systems development where the programmer designs mock-ups and asks the user to try them out and give feedback. The advantage of prototyping is that the users do not have to wait until the system has been fully developed, before they can try it out.

### 3.4. Standard operating procedures

SOPs are a set of written instructions detailing how a particular process is carried out. Computer-based systems support trial processes by providing a means of storing, modifying, and retrieving data. The process by which these computer systems are developed and used should be documented and controlled by procedures (SOPs) that ensure that they are adequate, and, where necessary, GCP-compliant. SOPs allow different people to check the procedures and ascertain whether what is done corresponds to what should be done. SOPs are also invaluable for training different people in the tasks that need to be undertaken in the study.

It is usually a good idea to split the SOPs into different categories such as database development, database validation and testing, database implementation and site set-up, and database maintenance/backups/upgrades. SOPs should be written by the person responsible for the task (who knows what should be done) and checked by the person who supervises their work and finally approved by the study PI.

### 3.5. Version control

The purpose of version control is to keep track of changes made to a computerized system during its development and after it has been implemented. The changes can come from various sources. For example, the users may find errors that need to be fixed when they start using the live system or may request new features or improvements to the system. Also, a change in environment may require a change in the computer system. For example, a decision to move from Microsoft Access databases to SQL Server databases will require changes to the data entry screens. Another example is a new compliance requirement that requires a certain type of report to be generated by the system.

A requirement for GCP-compliant data management is the use of validated and verified computerized data collection systems. Systems are validated and verified by thorough testing, comparing the database system against the user requirements. A validated and verified system is one that meets its specifications and requirements and that fulfils the purpose it was created for. Any change made to an already validated computer system may introduce new errors. Hence, the process of making changes needs to be done in a controlled environment. Previous versions of the program code are stored in a version control software, for example, Visual SourceSafe, Subversion. The system ought to be tested after making changes to ensure it remains in a validated state, before deploying the updated version. The detailed process for managing and maintaining changes to the system should be written in a version control and change management SOP. It is important that the SOP is adhered to strictly.

### 3.6. Confidentiality

Information that can be used to identify a person should be stored in a secure database that allows only authorized persons to access the data. Any data that can be used to identify a person, for example, name, address, date of birth, should be kept out of the public domain. Sensitive information should be identified from the onset, so that appropriate controls are put in the database. If the data are to be shared, it is necessary to decide how this will be done and what kind of security checks will be put in place. Technical security mechanisms, such as audit trails, access control using user logins and passwords, and permissions should be supplemented by data-sharing contracts and user training. Encryption should be used when sharing or carrying data on portable devices to ensure that unauthorized users cannot read the data, even if they get hold of the portable device.

### 3.7. Training

However basic the database system may seem, users should be adequately trained and should fully understand what they are doing. This training may be in the form of professional and in-house training and may involve using a prototype of the database in a pilot scheme. User training logs should be kept as evidence of training.

### 3.8. Pilot testing and database testing

User acceptance testing and pilot testing are commonly used to verify that the database performs well. In user acceptance testing, the end-users test the new database by entering data, following the SOP, and trying out the functionality provided by the database. The end-users feed back comments to the programmers and study leaders, who can make the necessary changes to the database programs and to the SOP that define how the procedures work. This is very useful, since database issues are identified early and rectified, before data have started being captured. Pilot testing also helps to identify potential issues that may arise when the study systems go ‘live’.

## 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 on to 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, openMRS). 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.

## 5. Managing data

Data management is a major task in most intervention trials. The main stages of the data management process are:

1. entering the data into a computer system
2. checking the data for errors and inconsistencies
3. organizing the data into an appropriate form for analysis
4. archiving the data.

Good data management requires a well-defined data management strategy—it is not something that will just happen. The complexity of the data management process will depend on the size and type of study. Attention to data management will greatly reduce the time needed in the analysis stage, because the data will be well organized and consistent and have fewer errors.

### 5.1. Data entry

The process of data entry will usually involve a data manager who designs the data entry screens, while other staff, such as data entry clerks or field staff, do the actual entering of the data. Creating data entry screens is not difficult but requires care. The data entry screen should follow the questionnaire, so automatic skips can be used to follow the skips in the questionnaire, and drop down menus can be used to show the same options as in the questionnaire for individual questions.

If data are being entered from paper-based forms, data entry can be double or single. Double data entry is routinely used to minimize typing errors and to ensure that the data in the database accurately reflect what was recorded on the forms. There are two main techniques used for double data entry. In one method, two data entry clerks independently enter the data without any knowledge of the other's work, and both entries are stored. A program, which may have to be written by the data manager, compares the two entries and identifies any discrepancies. The resolution of the discrepancies is generally referred to as 'verification' and can be done in different ways. In some systems, verification may involve both data entry clerks re-entering the specific data fields where discrepancies were identified and comparing the new entries. More commonly, a third person resolves the discrepancies, by referring to the original forms or questionnaires, and makes a decision as to the correct entry, changing the incorrect entry appropriately. Once the data have been double-entered and no discrepancies between the two entries are identified, the data are considered verified.

In the other method, which is used in specialist data entry programs, such as CSPro, the second person resolves the mismatches at the time of entry. After the first entry is complete, a second person enters the data, and any discrepancies are flagged up immediately, as the data are being entered. The second person must decide what the correct value should be and enter it accordingly. With this method, the second person is generally chosen to be more experienced and is expected to make the final entry, corresponding to what is on the form. This method is quicker but more prone to error than the first method described.

Facilities for double data entry are a feature in some software packages, for example, Epi-Info. In other packages, such as Access, double data entry must be set up when the database is created and would require considerable time and skill. One option is to use a package, such as Epi-Info, for data entry, before transferring the data to a separate package, such as Access, for storage and management.

Single entry of data is relatively rare and not recommended. It should only be considered if there are extensive checking routines, strong supporting processes, and technology in place to identify possible errors. Generally, the cost of doing a second entry of data is less than the costs of the additional data management required to clean the errors that that may remain after just a single entry of data.

The task of entering data should be conceptually separated from the task of analysis. Different software may be used for data entry, data checking, and analysis. The data entry system should be designed to make data entry as simple as possible. Simplifying the keying process will speed the task and make it less error-prone. Ideally, the data entry screens should closely resemble the paper form from which the data are being copied. Questionnaires should also be designed with data entry in mind. The data should be entered as recorded on the questionnaire. No hand calculations or transformation should be done before data entry—these can all be done during the analysis stage.

### 5.2. Data checks

Most data management time is taken up with checking the study data for errors and inconsistencies and 'cleaning' it. There are three main points to be considered when developing data checks: (1) deciding what will be checked; (2) working out when each check will be used; and (3) specifying how to resolve inconsistencies and errors identified by the checks.

Many software packages have inbuilt facilities for data checking. These automatic check programs can be set up when the database is created and run at different stages of the data management process. Before the check program is created, a specification document should be prepared, defining the data that will be checked and the errors that the program will be designed to catch. This document is usually written by the data manager or statistician, with input from the investigators, and is known as the data validation plan or check specification plan. Generally, the person who sets up the study database will also write the checking program and will be responsible for testing it. The program should be tested on 'dummy' data, before using it on the actual study data, to ensure that it is working correctly. Several 'dummy' questionnaires can be completed with deliberate errors and entered into the database; these can be used as test data for checking purposes.

Data checks can be incorporated into the data entry screens, so that illogical or implausible values are flagged up at the time of entry. Furthermore, the entry screens can be designed, so that they do not allow entry of invalid values, such as an impossible date or a value of '5' for a question that has '1' to '4' as the only possible answers. There are arguments for and against using data checks at the time of data entry. The checks will slow down the entry process, and, with double data entry, it can be argued that the checks are not needed to pick up errors in data entry, but they will pick up some data that have been incorrectly recorded on the questionnaire. If checks are incorporated into the entry screens, they should be designed so as to allow entry of invalid values, if that is what is recorded on the questionnaire. Otherwise, the questionnaire must be put to one side, until the error is resolved, and risks being lost or misplaced, so that it never gets entered in the database. Except for very small studies, it is often better to get the data entered in the database and to run checks to identify any errors afterwards, especially where there are a large number of questions that might need checking. However, interactive checking may be preferable when data are entered by the same field staff who completed the questionnaires earlier in the day. They may be slow typists but, having the interviews fresh in their minds, are more likely to be able to correct errors at the time of entry. However, in this case, it would probably be even better to consider having the data entered at the time of the interview, using electronic data capture methods.

After data entry is complete and the data are verified, an automatic check program is run to identify errors. These checks can include range checks to identify out-of-range or missing values (for example, dates out of the expected range, participant's age outside the range permitted by the study protocol) and cross-checks to identify inconsistencies between values (for example, males who are pregnant). The timing of when these checks are run requires careful consideration—for example, if a check compares data from different visits, the data from both visits must be present for the results to be meaningful.

In a longitudinal study, with repeated data collection visits to each subject, data checks should be run early on and continuously throughout the study. When errors are identified early in the study, it is often possible to uncover misunderstandings in the interpretation of the questionnaire or a flaw in the questionnaire design that was not picked up during the pilot phase. Clarification or further training can prevent those problems from recurring throughout the entire study.

The initial analyses are a continuation of the checking process and should include looking at cross-tabulations of the data to identify inconsistencies, and scatter plots and box-plots to compare groups and identify outlying observations. In large longitudinal trials, interim tabulations of data are recommended as a way of detecting possible data errors. Special checks might be made on observations that are more than two or three standard deviations from the mean. Such observations should be checked individually, as they are not impossible, merely unlikely.

Lastly, it should be noted that discrepancies in data are time-consuming to identify and resolve. The implications for data checks and query resolution should be considered during the questionnaire design stage. Asking for duplicate information in different parts of the questionnaire is one source of unnecessary queries. Queries will also arise if the questionnaire design does not make adequate provision for unavailable responses or permits ambiguous responses. Questionnaires frequently include questions that are deliberately used as cross-checks of other fields. This can be a very good policy when the data are actually different. For example, a check of sex against the subject's pregnancy status provides a reasonable cross-check of whether the person could not possibly be pregnant because they are male. However, problems arise when questions duplicate the same data, for example, a questionnaire that asks to record both the age and birth date. Discrepancies and confusion are bound to be generated when the values do not agree. When designing a questionnaire with requests for repeat information, consideration should be given to the implications for data checking and whether the duplicate information is truly needed.

## 5.3. Data cleaning

Data cleaning involves raising and resolving data queries that are identified during the data checking process and making the appropriate changes to the database to correct the errors. The aim is to be sure that the data are as of high quality as possible, before they are analysed.

Some data queries can be resolved within the data management group, for example, an obvious error in the year of a visit date. However, most problems will need to be resolved by the field team or the investigator. Commonly, the data management group will send a list of queries to the field team; the team will resolve the queries by writing the correct answer next to each one and return the list to the data management group. Alternatively, corrections can be made to the questionnaire itself. However, it is very important that the original answer is not obscured—instead, it should be crossed through with a single stroke, so that it is still legible, and the new information written on the side. It is good practice (and required for GCP) for the person making the change to initial and date the changes.

After the queries have been resolved, the database should be updated to reflect the corrected information. Some software systems have features to allow changes to the data after query resolution. These changes are usually made through the data entry screens and are recorded in an electronic audit trail. In a system without an automatic audit trail, changes may be made directly to the data tables themselves, although this can be more error-prone than changing the data via the data entry screens.

Correction or editing of data to reflect a resolution generally follows a different path from that of initial entry of the data. Most systems do not support double entry of corrections, so it is good practice to have a visual check of the data after correction to be sure that the change was made correctly. After the changes are made, it is essential to re-run the check program again, since it is possible that the update of the data has caused a new inconsistency to be identified.

In some cases, it may not be possible to obtain a resolution to a query, particularly if it is some time since the data were collected. Some software systems keep an electronic record of the problems identified by the check program, and a code can be entered to indicate that the inconsistency cannot be resolved. Alternatively, the incorrect data can be given a code for 'missing' if the correct answer cannot be obtained. The number of times this is done should be kept as small as possible and should be documented.

It is important to have a single master copy of the database that contains all the data corrections that are made. Even after the data cleaning stage is complete, errors may be detected much later during analysis; these should all be corrected on the master copy, so that it is always up to date. A version control system should be in place to ensure that it is possible to know which version of the database was used for any particular analysis.

## 5.4. Variable naming and coding

One of the first things to be done, before developing the study database, is to create an annotated questionnaire, containing the names that will be given to different variables and the characteristics of the variables such as numeric or text, the length of the variable (maximum number of characters), and any specific code lists that will be used. Ideally, variable names should provide information about the data being recorded, for example, 'birthwt' for birth weight or 'intdate' for interview date. In a longitudinal study, the same name should be used for those variables that are recorded at every visit. Some studies have used a convention of naming the variables at different visits with a number at the end corresponding with the visit, for example, 'visit_date3' and 'visit_date6' for visits 3 and 6. However, it is generally easier to run data checks and do other manipulations on the data if the variables have exactly the same name at each visit. An additional variable for the visit number should be included to identify the visit.

Some software packages have restrictions on the length of variable names; in particular, some older packages do not allow more than eight characters. A good general rule is to use no more characters than are allowed by the most restrictive (in terms of the number of characters) software package that is likely to be used for the study.

Questions that have categories of answers are best entered as coded values, rather than text (for example, 0 = No, 1 = Yes). These fields have a limited list of possible answers and only present a problem for data management if the field can contain more than one answer or if the answer falls outside the predefined list. When more than one answer is possible (for example, a list of types of contraception ever used), the database design changes from a single field to a series of fields, each of which can hold any of the valid responses from the list (coded yes or no). If an answer occurs that is not on the predefined list, a value for 'other' may be needed. In this case, it is advisable to create an additional database field where the specific response can also be entered as text. If an answer that is not on the list occurs frequently, it may be worth creating a new code for it. New codes may also be needed if the questionnaire design changes during the study or between survey rounds. In these cases, it is essential that the existing code list does not change. Instead, the new codes should be added at the end of the list. The coding of responses to questions is dealt with in more detail in Chapter 14.

Some studies use free-text on the questionnaire and re-code the text into categories at the time of data management. Re-coding variables in this way is generally not recommended, as, if they have been collected and entered, the original data should be found in the final data set. If any re-coding is done after data entry, the new data should be put into a new variable, with a note to indicate how the variable and the codes were defined.

### 5.5. Data lock

When all the data checks have been run, the queries resolved, and all QC activities are complete, the data are declared 'clean', and the database is 'locked'. This means that no further corrections will be made to the data. In GCP-compliant studies, the trial cannot be unblinded, i.e. the randomization codes made available to the study team and the main analyses cannot be performed, until the database is locked. At this stage, the locked database may be deposited with an independent body, such as the data safety and monitoring committee (DSMC), so that, if there are any later queries about the integrity of the data or changes that may have been made, comparison can be made with the locked set.

Even in studies that are not being run to be fully GCP-compliant, it is useful, at some stage, to make the formal decision that the data are 'closed', and no more corrections will be made. Sometimes, the data can be closed for one analysis, while corrections are ongoing for other data. The purpose of 'closing' the database is to ensure that the data are defined for a stable set of analyses and will not change every time the analysis program is re-run. The closing of the database, or any part of the database, should not be done before all the errors that are correctable have been resolved.

## 6. Archiving

New data are brought into a data management centre daily, and many different data changes and decisions are made. It is important that these are recorded and documented. If an accident happens (for example, a fire in the data centre), these changes and decisions could be lost and may be difficult to re-create, with potentially serious consequences for the integrity of the trial. This section advises on some of the ways to backup and keep the data, both for short-term protection and long-term use.

### 6.1. Interim backups

Backups of data are essential and should follow a regular pattern. Backups should not be thought of as an archive of the data, but only as a temporary store of the latest work. The procedure for backup should include times when a complete, full backup is made (perhaps monthly) and times when an incremental or partial backup is sufficient. The backup procedures should be documented in a SOP and agreed with the trial PI. Backups should be automatically scheduled, using a program or backup package, but one person in the study should be given responsibility to check the backup happens as scheduled. If the backup fails for some reason, that person needs to know what to do. At periodic intervals (preferably at least once per month), data should be backed up off-site, which can usually be easily and cheaply done onto an independent website.

What should be backed up? Everything should be kept in a backup, but not everything needs to be kept in every backup. The master database with the study data needs to be backed up regularly and completely. Other data that contribute to the master data should be backed up, and any changes recorded and backed up. Data entry files need to be backed up at least once but, as they should not be changed, may not need to be backed up again. Questionnaires and forms need to be included, as do coding sheets, reports, and correspondence with personnel inside and outside the study. Organization of the study data is important and should probably reflect the organization of the data on the main computer or server, and it should include a directory map to allow someone who is unfamiliar with the structure to find their way around.

An external hard disk is a cheap and easy way to make a backup. These are large enough to store many copies of the data (previous backups should not be deleted), but these external drives can suffer accidents and should not be considered a safe or secure storage of data. It is worth getting programs that will compress, encrypt, time-stamp, and validate the backed-up data to ensure that it does represent a true copy of the data at that time. Backups should not be considered a permanent solution, as technology moves on, and new systems and programs replace old ones. For example, backup data stored on floppy disks from 2000 were no longer readily accessible by computers or programs in 2012. This means that it may be necessary to copy backups onto new hardware/software every few years, before they become obsolete. And the final archived data sets must always be kept accessible on current hardware and software.

### 6.2. Metadata

An archive of the data is of limited use without the extra information that specifies exactly what the data comprise. These additional pieces of information are called metadata and can include information about the study setting, inclusion and exclusion criteria, the questions asked in any questionnaires, the codes for the variables, and a host of other information. Without such information, the data collected in the study are not interpretable. Note that metadata can include the names of the authorized users of the database and their passwords, as, without this information, it would not be possible to access the database and retrieve the data.

Extensible Markup Language (XML) is a set of rules that allow text, documents, codes, names, and even pictures to be stored in a machine-readable format. This allows the metadata for any study to be added to a repository and enhance the ability of others to use and understand the data. There are a number of XML schemes available, but, whichever is chosen, the metadata should be preserved for future use.

The Data Documentation Initiative (DDI) (see <http://www.ddialliance.org>) takes the storage of data and metadata one step further by defining a set of instructions for the storage, exchange, and preservation of statistical and social science data.

### 6.3. Data sharing policy

Usually, investigators will not allow sharing of the data from a trial with persons not directly involved in the trial, until the data collection and entry are complete, the trial has been analysed, and the main results published. However, at this stage, others may be interested in accessing the data to undertake further analyses or to combine the data with those from other trials to conduct a meta-analysis (see Chapter 3). Many funding agencies are moving towards insisting on sharing of data as a condition of funding. For example, the Wellcome Trust states that it is ‘committed to ensuring that the outputs of the research it funds, including research data, are managed and used in ways that maximize public benefit. Making research data widely available to the research community in a timely and responsible manner ensures that these data can be verified, built upon and used to advance knowledge and its application to generate improvements in health’. Most other major charitable or governmental funding agencies have a similar policy. The US Institute of Medicine published a consultation document in January 2014 on the guiding principles related to clinical trial data sharing (National Research Council, 2014), and their final recommendations in 2015 (National Research Council, 2015). Most large research institutions have a data sharing policy. The data sharing policy will define what data have been collected, stored, and will be made available, and the procedures to be followed for making some, or all, of the trial data available publicly or to selected recipients. Increasingly, the data collected in any trial, especially if it has been funded by a charitable or government agency, should not be thought of as belonging exclusively to the research team or to the director of the institute that conducted the trial but as a public good. After a reasonable period of exclusive access, it is widely accepted that the data should be made available to other researchers, policy makers, and medical authorities to further the advancement of knowledge.

The data sharing policy should be drafted at the start of a trial, as it will influence the way in which data are stored and archived. In particular, consideration must be given to how the strict confidentiality of the identity of the study participants can be preserved in any data that are shared. Furthermore, shared data are only useful if the recipient has a proper understanding of the information being shared. This requires that the data collection and coding systems are carefully documented for possible future onward transmission. This is one reason why metadata are essential.

### 6.4. Archiving hard copies

Paper copies of data and study procedures need to be kept for some time after the end of a trial. Some funders require these hard copies to be kept for periods in excess of 10 years after the completion of the trial, as the ultimate reference for the study data. Paper copies will need to be sorted and archived in a logical way. Space needs to be obtained for such storage, and protection ensured against fire, theft, and destruction by mould, insects, or other animals. Some studies are experimenting with scanning all documents and preserving the digital images instead of the hard copies, but this needs to be agreed in advance with the regulatory authorities and may not be acceptable to all. If data are collected electronically, the long-term storage of paper forms is no longer relevant. However, this puts even more emphasis on the need for careful and accessible archives of electronic databases, which should always include the original data as entered, as well as any final data sets.

## 7. Preparing data for analysis

The ‘raw materials’ for data analysis are the data files created by the data management process. However, the variables, as recorded in the questionnaire and entered into the database as raw data, are not always the ones directly suitable for data analysis. Re-coding and creating of new variables is likely to be necessary. It is generally also necessary to combine information from different data files.

When preparing the data for analysis, it is good practice to create a new data set with a different name to separate it from the original study data. Also, it is advisable to keep a copy of the commands used to prepare the data (either the program that was used or the ‘log’ files), in case it is necessary to re-create the file from the raw data.

### 7.1. Data dictionary

The data dictionary is part of the metadata and is the link between the questionnaire and the data files. It typically contains the name and a description of each variable, with additional information such as the data type (for example, numeric or text), coding (for example, 0 = No, 1 = Yes), and the questionnaire section and question number to which the variable relates. The data dictionary is essential for understanding how the data are structured and is used in preparing for data analysis.

### 7.2. Creating new variables

Sometimes, it is necessary to create a new variable from two or more existing variables, since this new variable may be more meaningful than the ones on which data were collected directly. For example, body mass index (BMI, defined as weight in kilograms/height in metres$^2$) or weight-for-age may be better markers of nutritional status than weight on its own. Such composite variables may be calculated directly from the raw data or be obtained by comparison with a given standard (as in the case of weight-for-age).

Variables related to time, such as the length of residence or the duration of exposure to a risk factor, present a special case. Depending on the characteristics of the variable and of the population under study, it may be preferable to record relevant dates on the questionnaires and to subtract them during the analysis stage to compute the duration of residence, exposure, etc. These calculations can be done, without difficulty, with any statistical package.

After creating a composite variable, it is useful to check that the distribution of the new variable seems reasonable. It is also appropriate to check the range of the new variable, as data errors may only show up at this stage. For example, negative ages or extreme weights-for-age may result from errors in the date of birth (or date of interview) in the questionnaire, though such errors should have been detected through consistency checks at an earlier stage.

### 7.3. Coding and re-coding

Before beginning the analysis, it is usually necessary to re-code some variables, so that they can be grouped into categories. Since it is advisable to look at cross-tabulations of data before moving on to regression methods, re-coding is generally needed for quantitative variables. Grouping makes it easier to understand the data and, in particular, to look for non-linear associations. But re-coding may also be necessary for categorical variables with large numbers of categories, or few observations in some categories.

When re-coding quantitative variables, one strategy is to divide the range of the variable into quartiles or quintiles, giving four or five groups with equal numbers of observations in each group. Alternatively, cut-off points may be chosen on the basis of established standards. For example, when grouping age, it is more natural to use 5- or 10-year age bands (for example, 20–29, 30–39, etc.), rather than base the categorization on quartiles. Similarly, there are recognized international cut-points for variables such as BMI (less than 18.5 is considered underweight) or weight-for-age (less than −2.0 is considered stunted). A histogram of the data is often a good way of deciding how to categorize a quantitative variable with no standard cut-points.

With categorical variables, it may be necessary to combine groups if there are very few observations in some groups. When combining groups, an important principle to remember is that, for combining to be appropriate, the risk of the outcome should be similar in each of the combined groups. For example, in a study of child malnutrition, it may not be appropriate to group mothers with no schooling with those with primary school education.

The number of groups to use also depends, in part, on how the variable will be used in the analysis. If the variable is an exposure of interest, where it is planned to examine the pattern of dependence of the outcome on the amount of exposure (for example, a dose–response), it is important to use enough groups to get a reasonable picture of the relationship. For example, to examine the effect of alcohol intake during pregnancy on birthweight, one group might be non-drinkers, and there could be four or five groups for different levels of alcohol intake.

After deciding if and how each variable should be grouped, the different categories should be assigned 'labels' to describe them. These labels should be saved in the data set, which will eliminate the need to return to the questionnaires or code lists during the analysis. When a variable is re-coded, it is important to create a new variable and allocate it a different name, so as to preserve the raw data. Thus, the variable 'AGE' might be grouped and allocated to another variable called 'AGEGP'.

### 7.4. Merging and linking data

The data required for a particular analysis may need to come from several different data sets (for example, questionnaire data on an individual's recent sexual behaviour may need to be linked to laboratory results, demographic data collected previously, and household-level data on the socio-economic status). If complete data tables are extracted for analysis, merging of the data may be more easily managed in the statistical package used for the analysis.

Many data management packages allow the construction of complex views of the data and can be used to extract merged data for analysis. The data analyst can specify the variables for analysis, and these can be extracted from the database, using standard data management tools, thereby maintaining the confidentiality of the data. It also enables simple data extraction programs to be used at regular intervals for longitudinal data, giving regular snapshots of the data for analysis.

## References

Hernandez, M. J. 2013. *Database design for mere mortals: a hands-on guide to relational database design*. London: Addison-Wesley.

McFadden, E. 2007. *Management of data in clinical trials*. Hoboken, NJ: John Wiley & Sons.10.1002/9780470181287 [CrossRef]

Murrell, P. 2009. *Introduction to data technologies*. Boca Raton: Chapman & Hall/CRC.10.1201/9781420065183 [CrossRef]

National Research Council. 2014. *Discussion framework for clinical trial data sharing: guiding principles, elements, and activities.*Washington, DC: The National Academies Press.

National Research Council. 2015. *Sharing clinical trial data: maximizing benefit, minimizing risk.*Washington, DC: The National Academies Press.

Powell, G. 2006. *Beginning database design*. Indianapolis: Wiley Publishing.

Prokscha, S. 2012. *Practical guide to clinical data management*. Boca Raton: CRC Press.

Pryor, G. 2012. *Managing research data*. London: Facet.

Bookshelf ID: NBK305509