20.6: 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 meta-data 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](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 shared and under what conditions.

https://med.libretexts.org/Bookshelves/Nursing/Book%3A_Field_Trials_of_Health_Interventions_-_A_Toolbox_(Smith_Morrow…

Updated: Fri, 09 Sep 2022 19:06:39 GMT
Powered by
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.