Clinical Data Management, 2nd Edition. Richard K. Rondel, Sheila A. Varley, Colin F. Webb
Copyright © 1993, 2000 John Wiley & Sons Ltd
ISBNs: 0471-98329-2 (Hardback); 0470-84636-4 (Electronic)

4 Data Capture

EMMA WATERFIELD

Clinical Trials Research Ltd, Maidenhead, Berks, UK

INTRODUCTION

The term 'data capture' refers to the accumulation of clinical data onto a database in a consistent, logical fashion so that it can be retrieved and searched. The content of the database should be an exact representation of the Investigator's observations at the clinical trial site, and capture of data must not obstruct this. For the purposes of this chapter, the term does not include the identification or interpretation of data errors or inconsistencies, except where this procedure is directly linked with the data capture step.

All companies, whether pharmaceutical/biotechnology organisations or Contract Research Organisations (CROs), will have similar objectives in mind when appraising the effectiveness of different data capture strategies. There is a universal need to submit data faster to regulatory authorities and also to submit to a larger number of authorities based in many more countries than was the case in the past. In addition, the volume of data collected in clinical trials has escalated, due both to the numbers of trials conducted and their increasing complexity and to increasing demands to prove drug safety. The chosen systems must therefore achieve the required balance between data quality and reduced drug cycle development time at the lowest overall cost.

Those involved in Clinical Data Management (CDM) are likely to experience an increasing pressure to review their data capture practices in an attempt to resolve this quality/speed/cost dilemma. By reviewing and understanding each data capture option, a considered judgement may be made as to which is most suitable for the individual organisation. Different solutions are likely to emerge depending on the individual factors involved.

This chapter explores the different types of data capture available to clinical data management, including the more established manual methods of data entry and newer electronic data capture technologies,

Clinical Data Management. Second Edition. Edited by R.K. Rondel, S.A. Varley and C.F. Webb.

the factors influencing choice of data capture method and some future prospects.

BACKGROUND

Historically, data capture methods have been restricted by available technology. The speed of computer systems, together with limited memory capability, restricted the numbers of users who could operate the system, the quantity of data and the complexity of programming possible. Reliability was also often a problem. Improvements to both hardware and software have developed to such an extent that technology is now rarely the limiting factor. Data capture will be considered firstly from this historical perspective, moving on to the newer technologies which in some cases completely bypass the methods utilised in the past.

TRADITIONAL DATA CAPTURE METHODS

Single/Double Data Entry

Traditionally, data capture has meant manual entry of data by trained specialist data entry operators who input data from a paper case record form (CRF) onto a central database via pre-set data entry screens, using a conventional keyboard. Data entry might occur only once (single entry) or successively (double entry), the latter with input by a second, separate data entry operator. Double entry aims to increase accuracy by highlighting the differences between the two operators' versions of the data. Reconciliation between the two entries may be achieved by either of two methods. On-line data point to data point verification by the second, more experienced data entry operator calls for a judgement to be made between the two conflicting entries, or flagging of the discrepancy for further investigation by CDM staff. An alternative is to run a report after double data entry which compares the two entries and indicates the inconsistencies. Subsequent comparison of the non-matching variables must then be accomplished by CDM staff before the data are transferred to the production area of the database. This represents an additional, time-consuming step in the data entry process.

The usual rationale for applying double data entry is that the increase in accuracy outweighs the additional expense and associated time delay in twice entering the data onto the system. One approach to maximise efficiency might be to employ single entry for text fields (which are notoriously difficult to enter accurately and are often subject to a later listings review by CDM staff), with double entry for non-text fields. In addition, by assessing the relative importance of different fields of the CRF to the final analysis, certain items of non-critical data might also be entered only once.

Legibility of text is likely to be a significant problem for the data entry operator. In general, operators are asked to identically reproduce the CRF page content onto the database, making no assumptions about the data that they see in front of them. In most cases, it is advantageous to use a flagging system, whereby illegible text is indicated by a keyed symbol which can later be reviewed by more medically aware CDM staff, thus reducing input time by the data entry operator and avoiding duplication of effort.

Data Entry Screen Design

Design of data entry screens is an important factor in determining the speed at which data can be entered. In general, the more similar the screen looks to the original CRF page, the easier it will be for an inexperienced data entry operator to enter the correct data in the appropriate field. However, experienced data entry operators often key very quickly, barely glancing at the data entry screen. If standard sets of CRF page templates are used to design the CRF, standard screen templates can be produced to minimise design effort for each individual study. Since similar pages are often repeated throughout the CRF, for example vital signs recorded at every visit, use of a single screen template for all such occurrences is often employed as a strategy. There is an additional benefit in maintaining a library of CRF page templates and matching data entry screen templates, since data entry operators will become familiar with the standard layout and require less study-specific training. A simple layout with individual data fields progressively one under the other is probably more effective when trained data entry operators are to perform the keying, as opposed to data fields placed randomly on the page. Obviously, ordering data entry screens in the same sequence that they occur in the CRF book also facilitates the entry of data.

It is often advantageous to program an ‘index’ table into the database which contains the unique patient identifiers, namely the protocol number, centre number and patient number. The index table can then be linked with all other tables in the database, allowing automatic population of these variables as each subsequent screen is accessed, thus simultaneously conserving data entry time and maintaining accuracy in key variables. A further benefit of such an index is that entry of duplicate records can be avoided.

Another consideration when designing data entry screens should be consistent formatting of analogous fields such as dates, which should preferably be entered in the same format throughout the database, for example always dd/mmm/yy in a single field or dd, mm, yy in three separate fields. The latter format may aid recording of partial dates, since a month and a year can be captured, even if the day is unknown. Attention should also be paid to the field attributes, (alphanumeric, integer, floating decimal, etc.) so that these are kept as consistent as possible across like fields, particularly where tables may later require merging, for example with an autoencoder program or imported Central Laboratory results, or be prepared for data transfer.

Restricting the input of data to a limited nomenclature can be achieved by the use of codelists. This reduces the possibility of error by the data entry operators, since only a finite number of keyed responses will be permitted, e.g. 'Y' or 'N' for 'yes' and 'no' respectively. Entry of any value other than those specified in the codelist would then be notified to the data entry operator, so that the correct value could be instated. As mentioned previously, ability to enter flags highlighting illegible or missing data can be a valuable facility, and should be designed to allow data entry operators to operate them using as few keystrokes as possible.

The draft data entry screens must be validated before use, both by the programmer who designed them and by the end user, preferably using test patient data. This provides assurance that the data typed into the screens is equal to that stored on the database, and assists creation of a user guide. A database listing of the test patient data should be checked against the original to confirm that the data are identical. All documentation of such validation should be signed, dated and retained in order to comply with GCP/ICH guidelines.

Centralised vs Local Data Entry

Data entry occurring in one central location is still standard practice for most companies. However, there is a time delay in mailing/couriering the paper CRFs to the central site from what may increasingly be multiple worldwide locations, as companies adopt globalisation policies. One solution would be to have entry occur at the Investigator sites or at several local office sites but this brings with it certain disadvantages:

- Entry staff less experienced with entry of data
- Requirement to maintain systems in remote locations and provide servicing and technical support
- Unlikely to be resource for second entry, therefore increased chance of errors
- Expense of initial outlay and ongoing support of data transfer mechanisms back to central site
- Compatibility of systems

By performing a company's entire requirement for data entry at a single central site, factors such as systems support, maintenance, stability and security can be more easily monitored and achieved. Trained dedicated staff are likely to be available on an ongoing basis to maximise the efficiency of the process. However, the nature of clinical research is such that flow of data in-house is unlikely ever to be at a constant rate, rather the organisation may at times experience a glut of data requiring urgent entry onto the database, whilst at other times it may be subject to periods of very low volumes of data requiring entry. An organisation may therefore find it practical to staff its data entry department with a baseline level of dedicated staff and supplement this with temporary staff as and when required. The use of temporary staff may not be an ideal solution to the problem of fluctuating data entry resource needs since the issues of training, quality standards and security must be considered.

An alternative solution to the challenge of retrieving paper CRFs from distant sites to either a central location or multiple local offices in the shortest possible timeframe is to use facsimile (fax) technology. This technology will be examined separately later in the chapter.

INCORPORATING DATA FROM AN EXTERNAL DATABASE

Manually entered paper CRFs may still be the norm in most companies engaging in CDM, however, many have embraced the opportunity to link with external databases to simplify the transfer of large volumes of specific data to their own database; for example, laboratory results from a Central Laboratory or blood pressure measurements direct from a monitor connected to a patient. Files of data can be downloaded onto disk, tape or via modem link and uploaded directly into the sponsor database, thus bypassing the need for a manual entry step. This is particularly advantageous in the case of laboratory data since before the availability of Central Laboratory data, entry proved very time consuming, especially when many different normal and alert ranges, units and repeat values had to be recorded.

Precautions must be taken to ensure that data integrity and security are maintained when data are transferred electronically. Either of two approaches may be selected: (i) ensuring that the two databases are compatible, or (ii) that suitable conversion programs are generated. Both systems must also be validated, common variables (e.g., protocol number, study number, centre number, subject number and initials, visit identifier) must be reconciled, and a procedure put in place for highlighting and resolving inconsistencies.

FAX-BASED DATA CAPTURE

Use of fax technology has accelerated throughout all industries in recent years. The potential benefits to CDM in speeding up the process by which clinical data can be gathered and tracked from diverse international sites and captured onto a centralised database have been quickly recognised, such that there are now a number of commercially available software packages designed specifically for use in CDM. The fact that the process can be relatively easily integrated into current working practices means that the idea of faxed CRF pages is more readily acceptable to those wary of, for example, remote data entry, since it represents a stepping stone between traditional paper-based data capture and full electronic data capture, requiring just a re-engineering of the paper CRF process.

Many sites will already have access to a fax machine, but if necessary, equipment can be provided at a relatively low cost and offering a high degree of resolution in terms of print quality. Maintenance and servicing costs must, however, also be considered. Provision of a (pre-programmed) free-phone number enables a user-friendly route to a central fax server, and training requirements are minimal. Transmission of data to the sponsor is less of a security concern using fax-based technology. Consideration must be given to the problem of tracking and reconciling duplicates of CRF pages of the type that are often updated periodically during the trial, for example adverse event pages or concomitant medication pages, which may have ongoing entries.

AUTOMATED DATA ACQUISITION FROM OPTICAL IMAGES

The drive to speed up the process by which clinical data are captured onto a centralised database has seen the development of more sophisticated scanning technology. When scanned, each mark on the original CRF page is viewed as a matrix of tiny dots which are stored electronically in the system so that an image can be assessed on a VDU screen rather than printed as a hard paper copy. The improvement over fax technology is that images can be recognised and stored in such a way that the information can later be deciphered. The step by which the optical image is interpreted onto the electronic database is known as automated data acquisition. Automated data acquisition can be subdivided into the following:

Optical Mark Recognition (OMR)

Where marks made on the CRF page in pre-determined areas are deciphered as meaningful data and converted to electronic values, e.g., yes/no check boxes on inclusion or exclusion criteria pages. Barcodes also rely on this technology.

Optical Character Recognition (OCR) Where the system can recognise numbers and characters. This is usually facilitated by restricting handwritten entry of individual digits or letters to specific boxes/areas.

Intelligent Character Recognition (ICR) A type of OCR system which has the ability to 'guess' at unrecognised symbols, retain a 'memory' of those previously encountered, and apply rules of association to enable interpretation.

Optical mark recognition can reach a very high level of accuracy if responses are restricted to check boxes. Optical character recognition as yet cannot match the levels of precision of OMR, but numerals recorded in boxes can be distinguished with a high degree of accuracy and short strings of characters such as patient initials are also fairly well identified. Accuracy levels for free handwritten text entries are poor, however. The degree of accuracy of data collected is obviously critical if subsequent automatic validation is to be performed effectively.

In order to make the transition to automated data acquisition, steps can be taken to modify familiar CDM procedures used for CRF design, database set-up and data entry screen design, thereby improving the likelihood of achieving an effective system. The designer will have to consider the ability of the scanning technology to comprehend the data, in addition to the more conventional requirements such as ease of data entry. It will be of paramount importance to maximise the ability to distinguish between marked and unmarked check boxes, and to identify the best methods of restricting handwriting in order to optimise accuracy of recognition. Another important technical requirement is the correct alignment of the image, which must be very precise in order for recognition to be performed. This can be facilitated by incorporation of location markers into the CRF page design.

Paper CRFs are collected from the trial site and either scanned at a regional office or central location, or faxed by the site directly to the main scanning point. Pre-printed study and page identifiers on the paper CRF are recognised by OCR and then subject and visit details can be used to uniquely catalogue the image and enable it to be tracked. A data entry reviewer can then check the data visually by on-screen comparison of image and data entry screen, and enter any unrecognised free text. Range and sense checks, including coding dictionaries and translation, can be programmed and the output reviewed by CDM staff, who annotate the screen with any queries before returning the image to the site for resolution. Queries annotated onto the CRF page tend to be more easily understood by site staff than traditional paper-based query reports. The process allows CDM to have earlier access to the data than is possible conventionally, and simplifies tracking and query management procedures. Storage of CRF images also greatly facilitates archiving obligations. Scanning data capture methods are effective for studies with numerous sites recruiting small numbers of patients, and for studies with complex data, which are more difficult to manage with the remote data entry process outlined below.

REMOTE DATA ENTRY (RDE)

Remote data entry means data capture at the point at which it is generated. If electronic transfer of the data to the sponsor's system occurs regularly, the data can be accessed and reviewed much more quickly than has been possible using traditional data capture methods. Crucially, the data can be validated as they are entered, by on-screen prompting, thus minimising errors early on in the process. The incentives for introducing remote data entry are therefore improved data quality, speed and flexibility of access to data, automatic avoidance of simple errors and early notification of error trends, which are all factors that can expedite time to database closure. Implementation of an RDE system is, however, very expensive. The following are likely to be significant costs:

- Purchase of a proprietary RDE package or design of an in-house RDE system
- Purchase of new hardware in-house
- Purchase of new hardware, e.g., PCs, laptops, palmtops, for remote sites and their delivery
- Communications charges and associated validation
- Training (both in-house and at site), support, e.g., helpdesk
- System maintenance and support
- Security of data transmission, integrity and access
- Compatibility with existing systems and data
- Re-engineering of processes and promotion in-house

Before investment in such a system, consideration must have been given both to the potential users of the system, and to the study requirements. Users will include Investigators, site staff, monitors, CDM, IS and potentially third parties such as a CRO or Central laboratory. Since a key element of the system is its integrated validation, all edit checks must be programmed in parallel with development of the electronic CRF (e-CRF). This makes it imperative that *all* members of the project team are involved in determination of the extent of data validation. Design of the screens must be from the perspective of the Investigator or site staff entering data, making use of the e-CRF as simple as possible, for example by logical sequencing of pages, on-screen prompts for missing or inconsistent data and skipping of irrelevant fields. There must be sufficient flexibility in the entry screens and edit checks to allow for data not available at the time of entry, which could also necessitate some checks being run separately in-house after transfer of data from the remote site.

The Investigator or designated site staff access the password-protected e-CRF on either a PC or laptop provided by the sponsor. By supplying a laptop, the sponsor can be certain that there is sufficient memory capacity and that the edit checks have been correctly loaded. Technical support of the trial may also be less complicated if all sites are using standardised equipment. Alternatively, use of a site-based PC would be cheaper and the software could be loaded and encrypted relatively easily. Data are entered onto the e-CRF directly from source documents which is aided by on-screen help messages and alerts such as range check outliers, protocol violation warnings or coding mismatches. The data are then downloaded at the end of each session to the sponsor's remote server via a modem link installed at the site. The sponsor, CRA or CRO can then immediately gain access to the data. Read-only, or read-edit restrictions can be placed as required to regulate access. Data management staff may initially be given read-only access to the data, and place electronic queries for any inconsistencies identified by their review, by either a pre-determined flag or using a query template. Once the Investigator has resolved the query to the CRA's satisfaction, the CRA can 'lock' the record, visit or patient data so that the Investigator no longer has edit access. This system has the advantage that the 'history' of a query can be scrutinised via the on-screen flags and date stamps, representing a dialogue between Investigator, CRA and CDM, and providing an in-built audit trail as stipulated by GCP/ICH guidelines. The e-CRF can be printed out once all checks are complete and returned to the site by the CRA at the final source data verification visit, where it is signed by the Investigator.

Some studies are inherently more suitable for RDE than others. It is important to assess the requirements of the study, for example the complexity of the data to be collected, the geographical location of the sites and the number of sites involved as well as training requirements. In general, RDE will prove to be most viable for simple studies with a large number of patients recruited at a small number of sites, particularly if the site is likely to be used by the sponsor company on a regular basis. Training and support of the site is also then more easily coordinated. Phase I studies lend themselves particularly well to the RDE process, since all the necessary elements are likely to be situated in a confined, easily regulated area.

Electronic Diaries

Data can also be captured remotely from electronic diaries supplied to patients utilising hand-held 'notepad' computer technology. Traditional paper-based patient diaries are notoriously inaccurate due to problems of patient compliance. Electronic diaries can be programmed to prompt patients to comply with the treatment regimen, and allow review of compliance since entry of data can be 'date-stamped'. Provided that diary data are downloaded at the end of each visit, assessment of compliance can be notified to the Investigator quickly enough that remedial action can be taken before the end of the study. This is of course dependent on the patients remembering to bring the electronic diary with them to the visit. Swift access to diary data also gives the sponsor earlier notification of adverse experiences. Training is an important precursor to effective use of electronic diaries. The Investigator must receive training from the CRA both in the operation of the equipment to capture the data and in the data transfer process which uploads the data from the diary. The Investigator must then be responsible for training the patient in the use of the equipment and in understanding the instructions. It is helpful if there is an in-built tutorial so that the patient can be tested in his or her understanding of the instructions prior to the start of the trial.

FACTORS INFLUENCING CHOICE OF DATA CAPTURE METHOD

The following factors might be considered when assessing the relative merits of the different data capture strategies.

1. *Initial costs.* Including outlay for new hardware and software, bespoke programming, training of Investigators, CRAs, CDM staff, process revision and documentation (new SOPs), validation of all new interfaces, security, organisational disruption.
2. *Ongoing costs.* Including training, support, e.g., provision of helpdesk, servicing, maintenance, communications, backups, ongoing validation.
3. *Accuracy.* Data quality, i.e., similarity of data to source data, ease of error identification.
4. *Speed.* Reduced time to database lock, lag times, set-up times, rapidity of entry onto database, accessibility of sponsor to data, potential for integrated validation.
5. *Security.* Patient anonymity and confidentiality, encryption, intellectual property/innovative ideas of sponsor, competitor interest.
6. *Flexibility.* Simplicity, adaptability to changing requirements/ environments (globalisation), compatibility with existing systems, reliability.
7. *Regulatory.* GCP/ICH requirements to be maintained, SOPs, audit trails.

The relative importance of each of these factors will of course depend on the individual company's goals, users, budget and study design. The traditional single option of manual input from a paper CRF onto a database has now been joined by technologies which have brought the ideal of 'point of generation' data capture to reality. These new technologies have wide-ranging potential advantages, specifically improved data quality, reduced in-house resourcing/ processing time and speedier access to data, but may prove prohibitively expensive in terms of hardware and software requirements, training and support costs, unless a true assessment of need has been determined.

FUTURE PROSPECTS

Remote Data Entry via the Internet

The potential of the Internet for use in CDM cannot be over-estimated. Set-up and access costs are cheap in comparison with the RDE methods mentioned above, and are becoming cheaper all the time. Many GPs and hospitals are already in possession of a suitable PC and telephone line and in theory it should be possible to give them access to the Internet at relatively little expense. The requirements would include a telephone line, modem and an Internet browser, which is the software used to read World Wide Web (WWW) pages. Of course in practice, the logistics are likely to be less simple. Security issues would need to be overcome by supplying encryption software, both to protect patient confidentiality and to preserve the sponsor's claim to original ideas and data. Testing of data entry and transfer mechanisms would also be mandatory for assurance of security. In addition, support and maintenance of the whole system would be difficult to administer if the PC belonged to the site and the rest of the hardware was loaned from the sponsor. It would be prudent of the sponsor to organise site audits of hardware and electronic transmission capability during the planning stages of a study, that is prior to the first patient being recruited and the associated need for immediate data capture. Suitability for the intended study could then be assessed after completion of a planned validation procedure, and preventive measures put in place to avoid potential corruption of data. An inventory of hardware would also be advisable, to document which components were the property of the sponsor. Resource would also be required for retrieval of equipment at the end of the study, unless a site maintains a regular association with the sponsor.

The Investigator and/or designated site staff would require password-protected accounts on the Sponsor's server, and would be able to access a specific website related to the clinical trial in which they were participating. The website might contain information about the trial, such as the protocol, recent amendments and advice, and the electronic CRF pages.

Clinical data could then be entered into the electronic CRF and pass directly to the sponsor's server. Simple checks on the data could be run in real-time, alerting the enterer to obvious range errors or inconsistencies, but more sophisticated edit checks may be less practical. A solution to this would be that more complicated cross-table checks could be run overnight, perhaps even utilising the processing power of the remote-based PC if edit check programs were downloaded from the sponsor's central server. Queries arising from the overnight validation process could be automatically electronically mailed to the site, alerting them to the fact that data entered the previous day required review. The conventional time delays of data entry and query turnaround would be drastically reduced, with the additional benefit that the Investigator would be addressing queries whilst the data were still fresh in his mind.

The great advantage of the Internet is thus the speed of communication possible. There would be a two-way benefit, for example the trial site could more quickly alert the sponsor of SAEs, whilst the sponsor could inform all sites, whatever their location, of new developments and instructions relating to the trial, thereby increasing the likelihood that any revised practices were implemented simultaneously. Compared with other methods of RDE, it is very quick and cheap to set up Internet-based systems, with the added value that interaction with standard software is both feasible and easy to validate.

There is less of a training issue with use of the Internet because the technology is simple and user friendly ('point and click') and therefore fast for novices to learn. This might be an important factor to consider if in the future an Investigator was involved in several trials for different sponsors, all utilising slightly different variations of RDE technology. The situation would be much more straightforward if all studies were managed on the Internet.

Direct Access to Medical Records

In the future, clinical research staff including those involved in CDM may be able to gain access to certain areas of GP or hospital databases, thus potentially negating the need for even electronic CRFs in certain circumstances. This would necessitate storage of medical records in a standard format for maximum effectiveness, and require encryption to totally anonymise each record. The issue of guaranteeing a patient's absolute right to anonymity is currently a highly debated topic and is likely to take some time to reach a resolution satisfactory to all concerned parties. There would be cost savings to the sponsor in that hardware would no longer be required at remote sites, there would be no ongoing need for support and maintenance and no training requirements for site staff.

Smart Cards

Another possible way of speeding up the capture of an individual patient's medical data would be to store these details on a 'smart card'. A patient's medical history, family medical history, details of past medications and previous participation in clinical trials could then be transferred automatically, and the card could be updated regularly with laboratory results, vital signs and changes to medication regimen. This idea again has implications with regard to patient anonymity and confidentiality, if the data were insufficiently protected.

Working from Home

The general trend towards working from home, whereby employees work at least part of their hours at home, has significant cost savings for the employer organisation and is made a step closer for CDM staff by the new technologies such as imaging, RDE and the Internet. We are likely to see increasing numbers of CDM staff able to complete their data management tasks from a home PC with modem link.

CONCLUSION

The scope of data capture requirements will vary widely between different companies engaging in CDM. For smaller concerns, traditional data entry from paper CRFs by data entry operators at a central location may still prove to be the most effective system when all factors are taken into account. However, larger companies which have pursued globalisation strategies and so benefited by the reduction in duration of their clinical trials, have established that data capture methods must be more efficient. As such, techniques such as remote data entry are becoming increasingly more widespread. The associated changes in the CDM process and ensuing restructuring of some elements of the organisation mean that the roles of those employed in CDM may become increasingly blurred with those of their colleagues in Clinical Monitoring and Application Development. The pace of development of technology is currently so rapid that there is the additional consideration for any company proposing to invest in new hardware and software of the hazard that it may become quickly out-of-date or redundant in a changing operational environment.