

A Review of Software for Data Management, Design and Analysis of Clinical Trials

B C Tai,*^{MSC, CStat}, J Seldrup,**^{PhD}

Abstract

In clinical trials, subjects are usually entered one at a time, and their responses to treatment monitored sequentially. Regular monitoring of trial progress in the early stages is crucial for accurate reporting of the final results. This paper discusses in detail the principles of quality data management in clinical trials, with specific reference to three clinical data management systems namely, CLINTRIAL, ORACLE CLINICAL and MACRO. All three systems have the essential features for monitoring and processing quality clinical trials data. In terms of functionality, there appears to be no significant advantage of one system over the other. However, to reap the full benefits of sophisticated systems such as these, a good support network and comprehensive training programmes are essential since their day-to-day use demands a high level of technical competence. Basic considerations pertinent to the success of a clinical trial involve not only logistics and data management. Issues relating to the study design are also of primary concern. In this respect, we briefly describe some sample size packages, NQUERY, PEST, POWER, POWER AND PRECISION and SAMPSIZE. Finally, a brief comparison is made with regards to some distinct features of three commonly used statistical packages, namely SPSS, SAS and STATA.

Ann Acad Med Singapore 2000; 29:576-81

Key words: Data management systems, Review of software, Sample size

Introduction

In clinical trials, subjects are usually entered one at a time, and their responses to treatment monitored sequentially. Regular monitoring of trial progress during the early stages is advisable, and prompt attention to data errors, inconsistencies or missing items on the case record forms (CRFs) is required, so that corrections can be made immediately. In the case of clinical trials of new treatments, for example, careful monitoring of patient safety is particularly important. If inadequate control is exercised in the management of clinical trials data, subsequent analysis may be delayed and/or results/conclusions inadequate or, in the worst case, wrong, because of erroneous or incomplete data. Hence, the need for efficient data management in clinical trials should be emphasised.

The use of computers for data processing and statistical analysis requires careful planning and execution by experienced personnel. Errors are liable to happen in the recording of data in the CRFs, as well as in the transfer of data to the computer using a database management software. Hence, data should be validated on-line or through other

data checking processes such as double data-entry which is still being carried out in some instances. Fortunately, with the advent of computers and specialised software for managing clinical trials data, the most tedious tasks of data processing and data checking have become more manageable.

This paper discusses in detail the principles of quality data management in clinical trials, and provides an overview of the general features of a good clinical data management system, with specific reference to three widely used software for managing clinical trials data, namely, CLINTRIAL,¹ ORACLE CLINICAL (<http://www.oracle.com>) and MACRO (Multimedia Applications in Clinical Research in Oncology, <http://www.infermed.com>). The paper finally discusses software for the design and analysis of clinical trials.

Principles of Quality Data Management

In a typical oncology trial, for example, the sequence of data collection might be depicted as shown in Figure 1.² Immediately following registration or randomisation, a

* Senior Medical Statistician

** Director

National Medical Research Council

Clinical Trials and Epidemiology Research Unit, Singapore

Address for Reprints: Ms Tai Bee Choo, National Medical Research Council, Clinical Trials & Epidemiology Research Unit, 10 College Road, Singapore 169851.

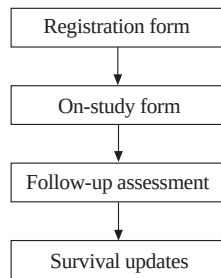


Fig. 1. Sequence of data collection procedure of a typical oncology clinical trial.

record of the patient's name, trial number, date of birth, date of randomisation and allocated treatment (if the trial is not blind) will be entered into the database. The on-study form which contains other demographic and baseline clinical information would be completed and added to the database in due course. Patients are then anticipated to return for regular follow-up assessments, say 6 weeks postoperatively, as in the case of a randomised trial comparing wide mesorectal excision versus total mesorectal excision, then 4-monthly for the first two years, half yearly for the subsequent two years and annually thereafter until death [National Medical Research Council (NMRC). *International Colorectal Cancer Trial: Randomised Trial of Wide Mesorectal Excision versus Total Mesorectal Excision for Mid Rectal Cancer*. Singapore: NMRC Clinical Trials and Epidemiology Research Unit, 1999]. Since patients do not all enter the trial at the same time, the

amount of data collected at any one time may vary considerably as the trial progresses, and the management of the data on a computer becomes complex and requires skillful intervention.

For repeated follow-up evaluations, a separate record is normally kept for each evaluation. For example, in a double-blind randomised placebo controlled trial on inoperable hepatocellular carcinoma (Asia-Pacific Hepatocellular Carcinoma Trials Group. *Randomised Trial of High-dose Tamoxifen versus Placebo for the Treatment of Inoperable Hepatocellular Carcinoma*. Singapore: NMRC Clinical Trials and Epidemiology Research Unit, 1998), information such as ECOG (Eastern Cooperative Oncology Group) scale of performance status, Okuda staging, serum albumin and bilirubin, as well as health-related quality of life information, are evaluated not only at baseline (on entry into the study), but also at monthly intervals. Since patients may drop out of a trial at any time for a variety of reasons (death, serious adverse events, lost to follow-up, etc), this further complicates the management of the database.

Quality data management via computers entails careful planning and execution by well-trained data managers. Figure 2 shows the sequence of a data management process of a typical clinical trial. As the processes of data transcription and entry into computers is highly susceptible to producing errors, a series of checks for validity and completeness should be carried out immediately upon the return of CRFs. Items which are commonly verified include

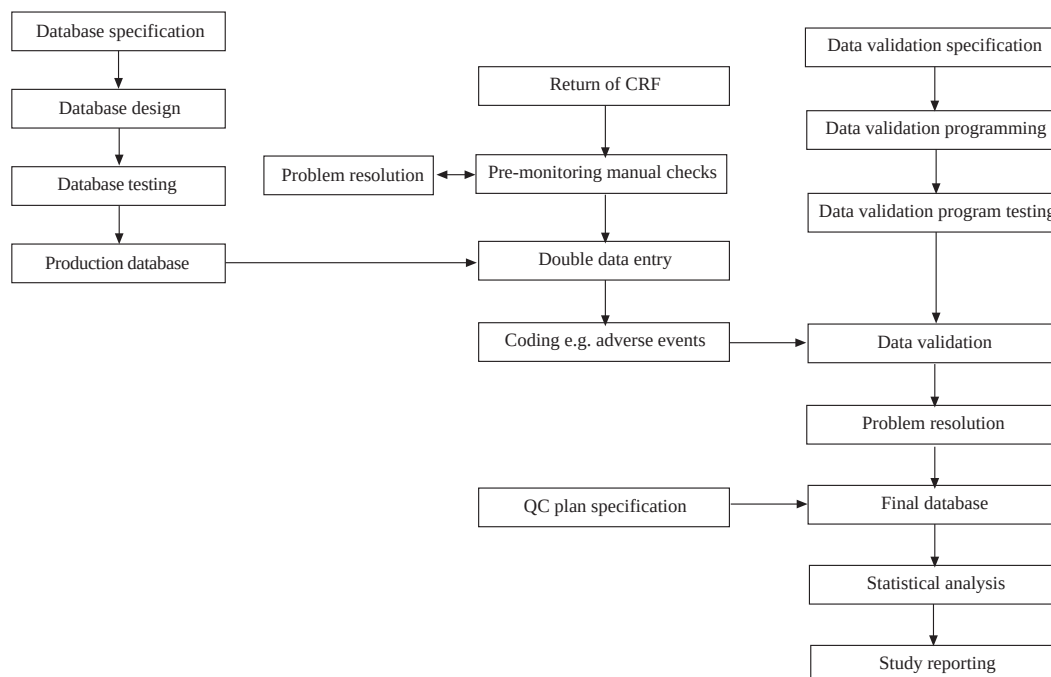


Fig. 2. Sequence of a data management process of a typical clinical trial.

range checks for clinical parameters such as serum levels which may be excessively high or extremely low (due to out of range or wrong units recorded), or in the case of qualitative variables, a non-permitted code such as 0 for gender when the CRF defines codes of 1 for male and 2 for female, respectively.

Logical checks identify any inconsistency in information which may be captured in different parts of the CRFs. For instance, dates of randomisation, follow-ups and death carry important information in clinical trials where survival is the endpoint. Thus, it is important to check that these dates as well as other crucial ones have been recorded and entered correctly. It is not sensible if the dates of death or follow-up of an individual is recorded as occurring before randomisation, and similarly the date of birth after randomisation. In addition, more subtle logical checks such as detecting large changes in clinical parameters, for example serum levels, between patient visits may also be implemented.

Routine checks on missing items or forms should also be undertaken, as any missing information could be due to oversight in data transcription, electronic data entry, or delay in returns of CRFs, rather than the information not being collected for various reasons. Queries should be raised for any discrepancies identified by these checks. The data should be edited accordingly after clarification with the investigator or comparison with source documents.

Although information in the CRF should be checked manually and discrepancies resolved by the data manager prior to data entry, on-line editing checks via computer provide an additional means of detecting errors or missing data. The validation rules would normally be specified and the associated data checks programmed in parallel with the building of the database.

Software for Clinical Data Management

Spreadsheets such as ACCESS,³ EXCEL,⁴ and even SPSS⁵ data editor are commonly used by clinicians for data entry of small clinical studies since they are relatively easy to use and readily available. In the case of the former two applications, no extra cost is incurred when using them since they are part of the Microsoft Office package generally available to clinicians. However, they are not suitable for managing clinical trials data which normally involve a higher level of complexity.

Several specialised software for managing clinical trials data are currently commercially available. The most common ones include CLINTRIAL, ORACLE CLINICAL and MACRO, all of which are Y2K compliant. Besides, these software are also developed to conform to the guidelines of Good Clinical Practice (GCP). Unlike

spreadsheets, all of them have the basic features of a good clinical data management system as described in detail below. Although other technical specifications and more cosmetic features are also important considerations, for the purpose of comparison, we shall concentrate on the essential data management features that have a direct impact on the quality of data processed.

An ideal clinical data management process is one that delivers valid and accurate data. CLINTRIAL, ORACLE CLINICAL and MACRO are data management systems which aid the maintaining of data integrity through the facilities for verifying and validating data, as well as through the implementation of an audit trail to document database modifications. They also allow for automatic reporting of discrepancies. For both CLINTRIAL and ORACLE CLINICAL, customised reports can be created and distributed to external sources such as the investigators for resolution. All three systems have the ability to efficiently handle repeated follow-up evaluations and track status of patients throughout the study.

These systems also provide global libraries to store the definitions of standard code lists, and standard validation or derivation criteria, to optimise information input and access throughout the operation. Standard codes such as 1 for male and 2 for female, as well as derived variables such as age, survival time, or Okuda staging and Child-Pugh score⁶ which are important prognostic factors in liver cancer trials, can be easily computed by defining appropriate derivation rules in the global libraries. These codes and rules can be tailored to meet the varying requirements of individual clinical studies and can be re-used in subsequent study databases. The testing of validation rules during the design stage of the database is allowed in CLINTRIAL and ORACLE CLINICAL.

A thesaurus management module is also available in such systems to assign a dictionary of standard codes to clinical data such as adverse events, drugs or disease. Some examples of coding thesauruses include COSTART (Coding Symbols for a Thesaurus of Adverse Reaction Terminology), WHODRL (World Health Organisation Drug Reference List), ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) and MEDDRA (Medical Dictionary for Drug Regulatory Affairs).

In multicentre trials involving several locations around the world, it is possible with these clinical data management systems to automatically propagate the study definitions, including amendments, to all study sites. At each site, the local data can then be independently managed using the same validation rules, derivation criteria and codelists. This is implemented by means of remote data entry, and the data can be incrementally replicated to any location for

query and analysis. Given the current trend for Internet access to clinical trials database and the use of faxed or scanning acquisition of CRFs, it is also possible for the faxed or scanned information to be exported into CLINTRIAL via an external software.

With multiple sites and many users accessing the same network, possibly performing different tasks, for example database design, entry and resolution by data managers, data retrieval, query and analysis by statisticians, the security of the system is of primary concern. The CLINTRIAL Admin module allows network-wide security standards to be enforced by enabling system administrators to assign, monitor and control access to sensitive clinical data records at the protocol as well as record level. Likewise, the security features of ORACLE CLINICAL include the ability to control user's access to specific study data or particular features in the system. Such feature is particularly useful in the management of treatment codes in a blinded study as individuals who are blinded with regards to treatment information may be restricted access to the specific panel where information on treatment is captured.

For the purpose of statistical analyses, there should also be fast, flexible and easy extraction of data into a variety of user-defined file formats, such as those of the more commonly used statistical packages. All three systems are designed for interoperability between the Oracle database and the SAS statistical package.⁷ For more advanced data query or extractions, users are required to be conversant in the Structured Query Language (SQL).⁸ However, more user-friendly robust query capabilities are also available in CLINTRIAL via SQL editors. These capabilities allow users to target and report relevant information quickly and easily.

Ideally, a good clinical data management system should have facilities for randomising treatments, and registering all information captured at randomisation directly into the database. ORACLE CLINICAL has a randomisation module, while developments are underway for both CLINTRIAL and MACRO. In the case of MACRO, however, links can be built with an external online randomisation system. As an example, the European Organisation for the Research and Treatment of Cancer (EORTC) uses MACRO for a randomised Phase III breast cancer trial on radiotherapy. In this trial, MACRO is linked with the EORTC's online randomisation system, ORTA, for assigning treatment. As the user enters demographic details, eligibility criteria and stratification factors, MACRO immediately validates the data and contacts the randomisation server to receive a unique identification number and treatment allocation, if the information entered is correct.

In terms of functionality, there appears to be no significant advantage of one system over the other. All three systems

have comprehensive features that meet the basic requirements of an efficient clinical database management system. However, a good support network and comprehensive training programs are vital since the day-to-day operation of these software demands a high level of technical competence.

In institutions where data managers or trial coordinators have limited exposure to computer tools and applications, in-house IT support is essential to reap the full benefits of a sophisticated system. Such assistance would facilitate the calculation of derived variables involving complex functions, the construction of advanced validation rules, and/or the application of SQL.

Software for the Design of Clinical Trials

Basic considerations pertinent to the success of a clinical trial involve not only logistics and data management. Issues relating to the study design are also of primary importance. A good clinical trial design is one that advances the answer to the scientific question of interest using the minimum number of subjects/patients. In a clinical trial, an excessively large sample size not only incurs higher cost but is also unethical by the involvement of "extra" subjects. Too small a trial size leads to inconclusive results, since there is a greater chance of missing the clinically important difference. The trial would thus be a waste of resources and is unethical.

Usually, the trial size is formally estimated using sample size packages such as NQUERY (<http://www.statsolusa.com>), PEST,⁹ POWER,¹⁰ POWER AND PRECISION¹¹ and SAMPSIZE.¹² The abbreviation PEST denotes Planning and Evaluation of Sequential Trials. As the name suggests, this software is developed for the design, simulation, monitoring and analysis of sequential clinical trials, including group sequential methods, the use of interim analyses and fixed sample size determination. Like PEST, SAMPSIZE is geared towards the design of clinical trials. In addition to the usual sample size calculation for Phase III designs, SAMPSIZE also allows for sample size estimation of post marketing surveillance and Phase II studies.

NQUERY, POWER AND PRECISION and SAMPSIZE are window-based and user-friendly, while POWER is DOS-based, and PEST is a SAS/AF¹³ application for Windows. Except for POWER AND PRECISION where only binomial and normal data types are available, all of the above software calculates sample size for outcomes involving binomial, normal and survival data types. Sample size for ordered categorical response types can also be determined using SAMPSIZE and PEST, and it is possible to obtain sample size estimates for agreement studies with SAMPSIZE and NQUERY. Graphical features for displaying the relationship between power and sample size

are also available in POWER AND PRECISION and NQUERY.

In most of these packages, for the purpose of hypothesis testing (difference or equivalence) users can calculate the sample size, power or effect size; for the purpose of estimation, sample size or confidence interval width may be determined. For most problems, trial size with equal or unequal number of patients per treatment group can be estimated.

Software for the Analysis of Clinical Trials Data

Statistical guidelines for contributors to medical journals¹⁴ and checklists for statistical review of papers for reputed international journals such as the British Medical Journal,¹⁵ now require confidence intervals to be given for the main results, instead of only presenting the *P* values from associated hypothesis tests. In commonly used statistical packages such as SPSS,⁵ SAS⁷ or STATA,¹⁶ confidence intervals can normally be obtained for the endpoint of interest. Even if the data have been managed and analysed by other means, confidence intervals can still be easily obtained using the Confidence Interval Analysis (CIA) software, which was written for use with the book "Statistics with Confidence" by Altman et al.¹⁷ This Windows-based software calculates confidence interval for outcomes involving binomial, normal and survival data types, and allows the user to input the raw data, or enter summary statistics, and define the level of confidence.

Commonly used statistical procedures can usually be found in standard statistical packages such as SAS, SPSS or STATA. More advanced statistical tools such as the cumulative logit for the analysis of an ordinal response variable, the logistic regression for a nominal response variable or the conditional logistic regression for matched case-control studies, can also be implemented in SAS or STATA. For more advanced statistical procedures that are not available in standard statistical software, macro programs can usually be written in SAS or STATA to implement them.

Other features that distinguish the packages are the quality of graphical output and the presentation of statistical tables or results. Of the three packages, SPSS produces the best graphs, and the editing of statistical output can be carried out quite easily. For academic clinical research where data is commonly captured using spreadsheets, data importation can be implemented easily in SPSS, although spreadsheets can also be directly imported in SAS. However, as the results of statistical analyses cannot be edited in SAS, it is hence the software of choice by many users including regulatory bodies.

There are also other specialised software that have been developed for specific types of analysis. The phenomenon

of missing data is commonly encountered in the assessment of quality-of-life in clinical trials due to death, lost to follow-up, or patient may be too sick to provide a response. Missing values lead to less efficient estimates because of the reduced size of the database. A software which offers principled approaches to missing data analysis is SOLAS (<http://www.statsolusa.com>).

Multilevel or hierarchically structured data is also not unusual in medical research. In clinical trials, several centres may be involved and the outcome of individual patients studied within each centre. Since it may be possible that centres differ in terms of patient care for example, the analysis should thus take into account the centre effect. While programs can be written using SAS or STATA to implement multilevel modelling,¹⁸ MLwiN¹⁹ is commercially available to specifically perform this task.

Conclusion

A well-designed study and quality data management play an important role in medical research, since they have a direct impact on the results of statistical analyses. A well-designed study with accurate and valid data, if poorly analysed, can be rescued by a re-analysis. However, a badly designed study with inferior data is beyond the redemption of the most sophisticated statistical technique. Thus, the successful conduct of a clinical trial requires the implementation of a good clinical data management system to monitor and process quality data in accordance to GCP guidelines, proper design of the trial with formal estimation of sample size, and efficient data analytical tools.

REFERENCES

1. Domain Pharma Corporation. Clintrial™ 4 release 4.2 Reference Guide. Massachusetts: Domain Pharma Corporation, 1999.
2. Pocock S J. Clinical Trials: A Practical Approach. Chichester: Wiley, 1993:160-75.
3. Jennings R. Special Edition Using Access 97. 2nd ed. Indianapolis: Que Corporation, 1997.
4. Chester T, Alden R H. Mastering Excel 97. 4th ed. California: Sybex Inc., 1997.
5. SPSS Inc. SPSS® Base 8.0 for Windows® User's Guide. Chicago: SPSS Inc., 1998.
6. Pugh R N H, Murray-Lyon I M, Dawson J L, Pietroni M C, William R. Transection of the oesophagus for bleeding oesophageal varices. *Br J Surg* 1973, 8:646-9.
7. SAS Institute Inc. SAS® Language and Procedures: Usage, Version 6. 1st ed. Cary, NC: SAS Institute Inc., 1989.
8. Dalton P. Microsoft SQL Server Black Book. Arizona: The Coriolis Group Inc., 1997.
9. Brunier H, Whitehead J. PEST Planning and Evaluation of Sequential Trials, Version 3, Operating Manual. Reading: MPS Research Unit, 1993.

10. Dupont W D, Plummer W D. Power and sample size calculations: a review and computer program. *Controlled Clin Trials* 1990; 11:116-28.
 11. Borenstein M, Rothstein H, Cohen J. Power and Precision: A computer program for statistical power analysis and confidence intervals. New Jersey: Biostat, 1997.
 12. Machin D, Campbell M, Fayers P, Pinol A. Sample Size Tables for Clinical Studies. 2nd ed. Oxford: Blackwell Science, 1997.
 13. SAS Institute Inc. SAS/AF® Usage and Reference, Version 6. 1st ed. Cary, NC: SAS Institute Inc., 1989.
 14. Altman D G, Gore S M, Gardner M J, Pocock S J. Statistical guidelines for contributors to medical journals. *BMJ* 1983; 286:1489-93.
 15. Gardner M J, Machin D, Campbell M J, Altman D G. Statistical checklists. In: Altman D G, Machin D, Bryant T N, Gardner M J, editors. *Statistics with Confidence*. 2nd ed. London: British Medical Journal, 2000:191-201.
 16. StataCorp. Stata Statistical Software: Release 5.0. College Station, Texas: Stata Corporation, 1997.
 17. Altman D G, Machin D, Bryant T N, Gardner M J, editors. *Statistics with Confidence*. 2nd ed. London: British Medical Journal, 2000.
 18. Goldstein H. Multilevel Statistical Models. 2nd ed. London: Arnold, 1998.
 19. Goldstein H, Rabash J, Plewis I, Draper D, Browne W, Yang M, et al. *A User's Guide to MLwiN*. London: Institute Of Education, 1998.
-