

Design Issues of Clinical Trial Data Management Systems

a report by

Antonio G Oliveira

General Director, DATAMEDICA Limited

The Clinical Trials Data Chain

Recent discussions on clinical trial data management have highlighted the concept of the data chain, a term used to help visualise data as a stream of information that flows across all phases of the clinical trials cycle, with data collected at each step in the process and used as a common resource by all clinical trials related activities, from protocol authoring to final reporting and publication.

In formal terms, the notion of data chain maps simply to a unified view of clinical trial data. Although information systems theory has recognised for a long time that all enterprise data should ideally be stored and managed by a single database system, in the clinical trials world that has not been the general practice. Existing commercial software products for clinical trials are closely tied to the specific functions they are required to perform, possibly because of the multiplicity of tasks involved in the conduct of clinical trials, each one performed by personnel with different expertise and skills. This situation leads to database proliferation and all its associated problems, particularly data redundancy and loss of data integrity. Data often cannot be effectively transported between applications, causing duplication of data entry, loss of efficiency and increased maintenance costs.

Current Approaches to Data Management

Data management is one of the multiple tasks needed for clinical trials and is concerned mainly with clinical study data entry, data validation and verification. At the surface, data management seems a perfectly demarcated process, starting with the design of a database able to hold the entire content of the case report form and ending with data cleaning and database closure. Available commercial systems have, in general, adopted that view of clinical trial data management and have developed database applications that assist users in defining the database structure, generate data entry screens and create database queries. Database design with those products, therefore, is mostly driven by the specific data content of the case report form.

Those products, however, have failed to acknowledge the relationships of data management with other activities conducted upstream and downstream of the clinical trials cycle. For example, the database scheme is heavily dependent on the study protocol, because the latter determines the study plan, patient allocation, most of the content of the case report form and the statistical methodology, among other things. Statistical methods also determine the database schema, as data needs to be structured in a form that is compatible with the requirements of a specific analysis. Database design must also consider the information requirements of other clinical trial tasks, such as site management and good clinical practice (GCP) monitoring, in order to be able to provide the data needed by those tasks.

Back to Design Principles

According to good database design practices, the ideal approach to clinical trial data management includes the development of integrated clinical trials information systems (CTIS) supporting all data-related tasks. It is certainly not possible to build an information system from scratch for each new project. Therefore, it will be necessary to think in terms of enterprise-wide information systems and, consequently, in CTIS based on generic clinical trials data models.

The main issue, then, is whether it is feasible to find database representations able to hold the data from any possible type of clinical trial. Clinical trials are widely different in scope and aim, experimental design, study plan and collected data, and there is the fear that generic data models may not achieve the granularity required to capture all the small details of a given clinical trial. Yet, it must be recognised that there are communalities across clinical trials that can be explored to create common data structures reusable by different clinical trials and, ultimately, to achieve generic representations of clinical trial data. Recent regulatory efforts, most notably the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)-E3 guideline, made a huge contribution to the identification and definition of many data elements that are common across all kinds of clinical trials.



A Paradigm Change

At DATAMEDICA, we have approached the development of a CTIS precisely by exploring those communalities. Actually, only a small part of the clinical trials data is unique in each study and we have shown that a common representation of that data can still be obtained with entity attribute value tables, flexible data structures that can represent accurately most types of information. Eventually, we have achieved a comprehensive and semantically rich data model that is able to capture all of the data from virtually every kind of clinical trial, whatever its design or data content. This has been repeatedly proven by over 100 clinical trials that were managed by our system in its five years of operation.

Basically, DATAMEDICA's approach required a paradigm shift from the data-centric approach used by existing systems, to a new study-centric approach that is process-oriented and generic, in the sense that it accommodates both the data management and data flows required by any clinical trial. There are many advantages to this approach, some of the more significant from the end-user's perspective being:

1. Single database – a single database with a standardised user interface, managing all studies in a CTIS, offers lower maintenance and reduced learning time. A single database also eliminates the problem of integrating data from different studies and allows the natural creation of a repository of high-quality clinical data for further post-study analyses.
2. Data integrity – with a generic clinical trials data model, many of the required validation checks can be encoded in the database design as business rules, with the advantage of leaving most of the data validation and integrity checking to the database management system. This provides much tighter control over data integrity than is seen in conventional CTIS, where data checks must be defined and manually coded every time a new project is started (although some provide the facility of reusing previously coded checks).
3. Automated reporting – contrarily to data-centric systems, it is affordable to write highly complex queries since they are reused in all studies. The study-centric approach supports the development of powerful automated reporting systems, capable of creating the complex tables required by the US Food and Drug Administration (FDA) and ICH guidelines.
4. Knowledge-based – because a considerable body of methodological and statistical knowledge is embedded in the data model, this approach

supports, among other things, the development of expert analytical tools and eliminates the need for moving the data between the CTIS and the statistical software, currently the process used by commercial CTIS.

Benefits of Generic Models

The first two points highlight the benefits of modern concepts of good database design, but are particularly important as far as CTIS are concerned. First, they afford the possibility of reusing the same data structures for each new clinical trial (e.g. concomitant medication, laboratory data, adverse events), thus reducing dramatically the time needed to customise the system for each new clinical trial. Secondly, clinical trials make extensive use of clinical ontologies and only through a central database can mappings across competing ontologies be implemented.

The last two points provide the most striking argument in favour of generic CTIS. Statistical analysis and reporting is a formidable task, requiring the generation of hundreds of tables and data listings and over 1,000 calculations of significance levels, difference estimates and confidence limits. Typically, several database and statistical programmers and biostatisticians are needed to perform this task over a period of time that is measured in months. However, if the clinical safety laboratory data, for example, is stored in the same data structures for all clinical trials, it is possible to build an automated reporting system that will create all the tables and listings required for the final statistical report in a matter of minutes. The same applies to the other data tabulations, such as adverse events, vital signs, concomitant medication, drug accountability and so forth. Clearly, such a system would be able to produce most of the tables of a statistical report in a matter of minutes.

Knowledge-based Capabilities

The last point noted above, the possibility of embedding methodological and statistical knowledge into the data model, is crucial for the development of automated reporting. Only a knowledge-driven system is able to offer this functionality because an automatic reporting system needs to be able to make decisions regarding the data to be used in each tabulation. For example, the system needs to be able to decide which patients and observations are valid for a given analysis, as safety data is reported for a population different from the efficacy population, and there are different criteria for the selection of the several populations used for primary and secondary efficacy analyses. A knowledge-based system will also ensure, within reasonable limits, that the results yielded by a user query will have statistical validity.

Intelligent Data Analysis

Although a database management system can provide all of the data management and reporting functions mentioned above, it does not have the built-in advanced statistical functions needed to create tables containing results of statistical analyses. This functionality, however, can be incorporated into the system by developing an additional software component based on a professional statistical package, such as SAS® or STATA®. The data can then be transported out of the central database into that module, processed with a statistical analysis algorithm and then returned back to the reporting system for final table generation.

The COATI Clinical Trials Information System

Such was the solution adopted by DATAMEDICA for the development of Control, Assessment and Tracking of Investigations (COATI), the CTIS we use for all our operational clinical trial data. We created a rule-based system that is able to select – given the characteristics of the study design and type of baseline and efficacy variables used – the statistical methods adequate for each analysis and the results that need to be reported. This application can handle all but the most exotic statistical methods used in clinical trials analyses for nominal, multi-valued, ordinal, interval and survival data in one-, two- and multiple-arm trials, crossover studies and factorial designs, with or without stratification. It can automatically select baseline variables for inclusion as co-variables, conduct post-stratification analyses and subgroup analyses. It handles imputation of missing data, scale transformations and regrouping of study centres. This application uses STATA as a statistical server, but other packages could be used as well.

DATAMEDICA's CTIS is currently made up of a central database and four components: clinical trial management, GCP monitoring, report generation and intelligent data analysis. It has a client-server architecture based on a Microsoft® Structured Query Language server relational database engine back-end, and runs on Windows® and Linux platforms.

Experience and Results

The efficiency of COATI is not comparable with any existing standard or metric. As an example, setting up the database for a new clinical trial is performed by a biostatistician in less than 30 minutes, while in commercial systems this task usually needs to be performed by a database programmer and typically takes several days to a week to complete. For a clinical trial of average complexity, a fully ICH-compliant statistical analysis report takes less than two person days to produce, while conventional approaches require something in the range of three to six months to

complete this task. Additionally, and based on a comparison of activity reports made before and after the utilisation of the system, we verified that time savings in the creation of monitoring reports amount up to 70%.

In conclusion, the notion of the clinical trial data chain is appealing by the synthesis it makes of the entire problematic of clinical trial data management, but it demands an entirely new paradigm in CTIS design that has little in common with current approaches.

Future Developments

We believe that we are already well into this new paradigm and we know from experience that the promises it holds are true. However, we also have realised that we are still a long way from the ideal CTIS, mostly because new challenges are constantly being discovered. For example, we have realised that the two-tiered architecture of our CTIS is limitative of future developments of new software modules addressing other parts of the clinical trials cycle. The adoption of multi-tiered architectures will not solve the problem either. Therefore, we are now in the process of re-engineering our CTIS in order to adopt an Internet-centric, decentralised architecture, comprising loosely coupled components and based on eXtensible Mark-up Language and Simple Object Access Protocol technology to provide seamless data interchange and interoperability among modules.

We believe that such complex software architecture is crucial for the creation of a development environment able to support the unimpaired development of the large number of software components needed for total clinical trial data management.

Conclusion

The design of CTIS is a difficult and exciting area in clinical informatics. Unfortunately, regardless of the enormous promise it holds for speeding up the drug development process and for research enhancement, it has not received as much attention from the research community as it deserves. Hopefully, the situation will change as more people begin to realise the formidable challenges it poses in terms of systems design and software architectures. ■

Contact Information

DATAMEDICA Limited
 R. Rosa Araujo 34, 5th Floor
 1250-195 Lisboa, Portugal
 Tel: (351) 213 182 580
 Fax: (351) 213 142 281
 e-Mail: info@datamedica.pt
<http://www.datamedica.pt>