Review Article

Construction of databases: advances and significance in clinical research

Erping Long*, Bingjie Huang*, Liming Wang, Xiaoyu Lin, Haotian Lin

The State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China

Contributions: (I) Conception and design: H Lin; (II) Administrative support: H Lin, L Wang; (III) Provision of study materials or patients: B Huang, X Lin; (IV) Collection and assembly of data: E Long, B Huang; (V) Data analysis and interpretation: H Lin, E Long; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*These authors contributed equally to this work.

Correspondence to: Haotian Lin. The State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060, China. Email: gddlht@aliyun.com.

Abstract: Widely used in clinical research, the database is a new type of data management automation technology and the most efficient tool for data management. In this article, we first explain some basic concepts, such as the definition, classification, and establishment of databases. Afterward, the workflow for establishing databases, inputting data, verifying data, and managing databases is presented. Meanwhile, by discussing the application of databases in clinical research, we illuminate the important role of databases in clinical research practice. Lastly, we introduce the reanalysis of randomized controlled trials (RCTs) and cloud computing techniques, showing the most recent advancements of databases in clinical research.

Keywords: Database; clinical research, data management; reanalysis of randomized controlled trials (reanalysis of RCTs); cloud computing

Submitted Oct 05, 2015. Accepted for publication Nov 24, 2015.
doi: 10.3978/j.issn.1000-4432.2015.11.13
View this article at: http://dx.doi.org/10.3978/j.issn.1000-4432.2015.11.13

Introduction

Database technology, created in the 1960s, is defined as data collection associated with special structural organization, storage, and application in a computer system for certain purposes, and it is the most efficient tool for data management. Throughout the research process, data management is critical for clinical trials. As database technology is widely used in clinical research, it not only makes collecting and storing the data more convenient and reliable but also makes data analysis more efficient, greatly improving the efficiency of research and the rapid development of clinical studies. However, with the development of modern science and technology, the reanalysis of randomized controlled trials (RCTs) and cloud computing techniques have emerged in database applications, which have brought about not only challenges and opportunities but also more convenience and possibilities. Holistically grasping the important role and the latest developments of database construction in clinical research has become one of the most important daily tasks for the majority of clinical researchers.

Databases for clinical research data management, definitions and types of database, and key factors for their construction

Clinical research data management includes verifying the completeness and correctness of the data collection in clinical trials, ensuring the support of experimental data by statistical analyses, and illustrating and interpreting the results of experiments, all throughout the clinical research process. Traditionally, clinical data management has an internationally recognized general procedure (1), as summarized in *Figure 1*. Currently, clinical trial data acquisition methods are mainly classified as one of two

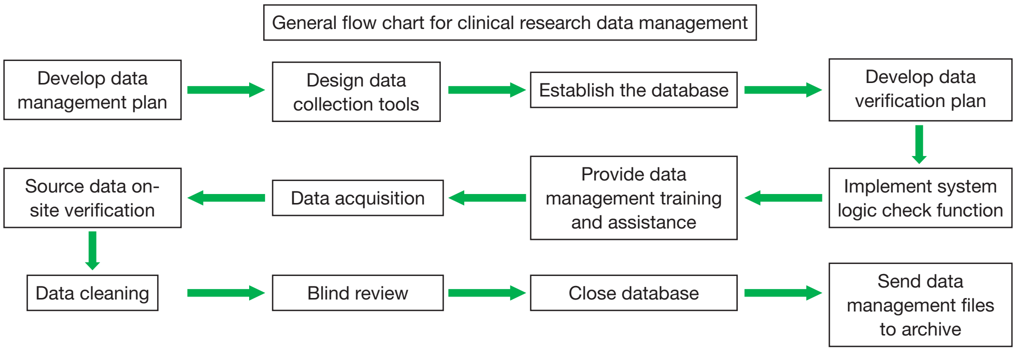
General flow chart for clinical research data management
Develop data management plan
Design data collection tools
Establish the database
Develop data verification plan
Implement system logic check function
Provide data management training and assistance
Data acquisition
Source data on-site verification
Data cleaning
Blind review
Close database
Send data management files to archive

Figure 1 General flowchart for clinical research data management.

types, i.e., paper data capture (PDC) and electronic data capture (EDC). Overall, EDC has the advantages of improving test efficiency and data quality, reducing costs, shortening the study period, decreasing loss, and enhancing the user's experience; the development and application of EDC tools in the field of clinical research have greatly promoted the process of digitalization in clinical data management.

A database is an advanced stage of data management and distinctively different from the traditional data management approaches mainly in the following aspects: first, databases make the data independent of applications, thereby enabling centralized data management and improving the efficiency of data through data sharing and reducing data redundancy; second, they establish links between different databases so that the connection between real-world information is reflected; therefore, the database is no longer just a simple tool for data collection (2). According to its nature, a database can be divided into one of three categories: literature databases, numerical databases, and factual databases, each of which can be further subdivided. Literature databases include professional literature databases, library bibliographic databases, combined catalog databases, patent literature databases, full-text databases, and others. Numeric databases include scientific databases, engineering and technology databases, and others. Factual databases include economic and business databases, management databases, public services databases, and others (3). All three main categories of databases have been widely used in clinical studies.

Based on the complexity of the data model, database technology can be accordingly divided into three stages, i.e., the first-generation nest and hierarchical database system, the second-generation relational database system, and the third-generation database system mainly featuring object-oriented models. With the increasing data types and complexity of the data association, nested and layered databases solely relying on the tree-root node styled structure have been unable to meet the growing demand for database functions and have been mostly discarded. At present, the most popular structure is the second-generation relational database, consisting of relational data structures. This type of database classifies, merges, and selects data through transforming the data structure into a two-dimension relational form and builds data platforms with complex and comprehensive management functions. However, different applications have specific functional requirements for databases that cannot be completely covered by relational databases. Therefore, the third-generation databases, with richer data models and more personalized functions of management and analysis, have emerged but have only been applied to a small number of fields, such as military applications, because of the limitation of their high cost (2).

Three factors are critical to successful database construction: management, data, and technology. First, management is the foundation of database construction, which includes decisions in the early stages and the formulation of corresponding policies, strategies, and measures; software development management, data acquisition, and quality control surrounding software engineering in the development stage; the security, confidentiality, updating, and restructuring of data; and the three types of maintenance management for application software (corrective maintenance, adaptive maintenance, and improvement maintenance) in the stages of service and maintenance. Second, data are the basis of database construction, and the reliability and stability of the data source are the prerequisites for everything else associated with the database, so stricter requirements for data

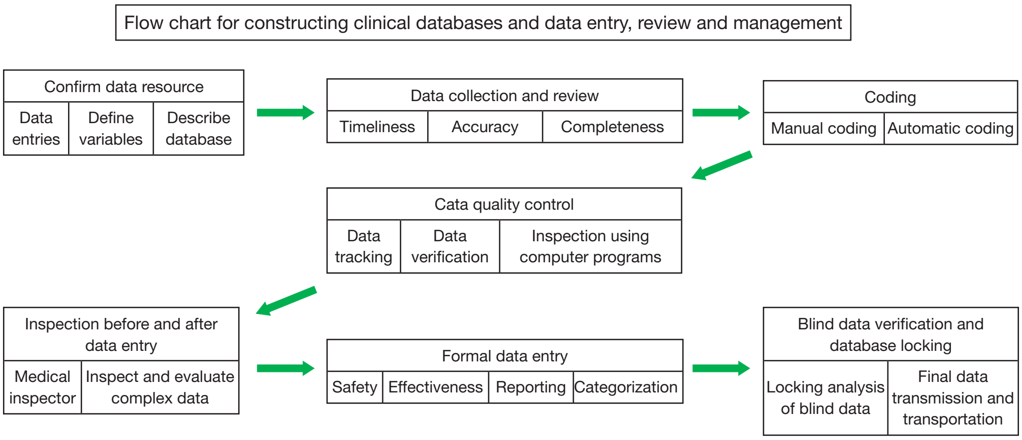
Flow chart for constructing clinical databases and data entry, review and management
Confirm data resource
Data entries
Define variables
Describe database
Data collection and review
Timeliness
Accuracy
Completeness
Coding
Manual coding
Automatic coding
Cata quality control
Data tracking
Data verification
Inspection using computer programs
Inspection before and after data entry
Medical inspector
Inspect and evaluate complex data
Formal data entry
Safety
Effectiveness
Reporting
Categorization
Blind data verification and database locking
Locking analysis of blind data
Final data transmission and transportation

Figure 2 Flowchart for constructing clinical databases and data entry, review and management.

collection, entry, and processing have been proposed. Lastly, technology is the most direct security for database construction and has wide and complex applications, mainly including database technology, computer technology, and network and communication technologies (4).

Construction of clinical databases and procedures for data entry, review, and management

Database construction, data entry, review, and management are very complex and meticulous processes, as shown in *Figure 2*. However, in actual operations, attention should be paid to the following aspects: (I) verifying the source of data, which mainly includes data from the site of the research and the associated laboratories. Then, the data to be collected (i.e., the research variables), the definitions of the variables (i.e., definitions for concept and operation), the description of the database (i.e., the data type, single or replicate), a confirmatory note on the consistency between the data and the raw data, and a confirmatory note on data validation (i.e., inspections on the values and the data range, missing values, and logic checks) are all entered; (II) collecting and reviewing the data, which requires timeliness, completeness, and accuracy. Specifically, first, "timeliness" is to ensure the quality of the raw data and to reduce the time required for the subsequent review; "completeness" is to require the collection of all the data on the subject; and "accuracy" is to require that first, the designed case report should have better operability, quantitative indicators should be adopted as much as possible, and appropriate quantification should also be applied to the soft indicators. Second, the laboratory conditions and operators should be kept relatively constant. Third, the clinicians or data collectors who collect the data or fill out the case report forms must be subjected to certain training. Fourth, recoding data from the raw data should be minimized; in case of recoding, careful reviews must be performed to ensure consistency between the case report forms and the original data. After completing the data collection, data review is necessary, which includes self-review and supervision and inspection; (III) coding, which includes manual coding and automatic coding techniques including dictionaries (MedDRA, WHO-ART, ICD, COSTART); (IV) data quality control. Commonly used methods include data tracking, data entry and validation, data verification, data questionnaires, checks using computer programs, and others; (V) data verification before and after data entry. Clinical data are verified by qualified medical inspectors; verification includes examining and assessing complex clinical data to find subtle difference in the data, after which data managers perform quality control on the evaluation; (VI) formal data entry and database construction. The requirements for data entry are either that problematic data should be resolved and the entry and verification of major data should be complete or the data have been sampled for quality control, and the major data related to safety and effectiveness have been evaluated with quality control checks, and all of the events that are not in accordance with research programs have been reported, categorized, and clarified with the related effects. After the entry has been completed and has met the requirements, data collation can be performed, and the data can be transmitted to the final database; (VII) blind data verification and database locking. After confirming the authenticity of the information and completing the blind data verification, the database is locked, and the process proceeds to the final analysis; the objectives and order of the analysis are further clarified together with the researchers. Then, the database is accepted, and a preliminary analysis on the information is conducted, which is followed by a preliminary data analysis to develop the draft report for the last analysis or revised statistical analysis plan. After verification, the database should be locked to prevent misuse and unauthorized modification, and a list of data required by the clinical study report should be generated. Lastly, the data are transmitted and archived, in which the data are sent to sponsors, statisticians, and compliance supervising departments (5).

The important role for the clinical database on determining the level of clinical research

The important role for the clinical database on determining the level of clinical research is mainly reflected in the following aspects: first, a clinical database is a prerequisite for data analysis and reaching a conclusion; second, a clinical database is one of the measurement bases for the level of research. The huge amount of data generated in a clinical study can greatly improve both the efficiency of research and the value of data through regularizing the database, and it can provide lessons and bases for researchers in the long-term process of the study. Analyzing the database can test the clinical study and reveal significance as well as determine the authenticity and reliability of the clinical study, thus effectively regulating the design and implementation of the clinical study (6). The data in clinical research are collected and managed in real-time, and any data management mistakes may break the original tightly interlocking links in the clinical trial, resulting in a loss that cannot be reversed afterward. The quality of the data management reflects the implementation status of the experimental design and scheme by the investigator in the clinical trial, while alluding to the scientific approach and knowledge level of the clinical researchers. Data management is also often subjected to the largest part of the audit of the implementation of clinical research and is one of the major responsibilities of supervisors and inspectors (7).

Effect of the reanalysis of RCTs on the conclusions of a study

The reanalysis of RCTs is one of the latest developments for databases applied to clinical research. In 2014, the Journal of the American Medical Association published a landmark study on the reanalysis of RCTs, in which the investigators found that in 37 cases, only five authors who reanalyzed RCTs were totally unrelated to the initial experiments. In terms of the experimental method, in the 37 cases, 46 different methods, including statistics and analysis methods, were applied, and in the RCT reanalysis results, 35% of the reanalysis drew different conclusions about patient treatment method from the initial experiments (8).

This experiment representatively indicated that the reanalysis of RCTs on data from existing clinical studies is feasible and can come to either the same or different conclusions as the original trial. The role of the reanalysis is to examine and supplement the initial experimental results or to find loopholes and errors. However, this analysis method still has some risks that have sparked disputes in academia because of its short history. First, the reanalysis of RCTs may pose a potential safety threat to patient privacy, resulting in the exposure of patient medical information. Second, inappropriate mining on data sets has occurred in the reanalysis of RCTs. Third, forged and false results have emerged in some reanalysis of RCTs. Fourth, some reanalysis of RCTs led to the leak of trade secrets (9).

Currently, only a limited number of reanalysis of RCTs have been reported in the literature, and the investigators of the reanalysis and the methods that were adopted all affected the results of the analysis (8,9). Therefore, at present, the quality and credibility of the reanalysis of RCTs are worrisome, and it is imperative to promulgate standards of data sharing and RCT reanalysis that are more authoritative and have a broader influence. Only after a unified and authoritative standard is developed will it be possible to mitigate various issues such as the violation of patient privacy, trade secret leaks, or experiments performed to advance a particular agenda, among others, so that RCT reanalysis can be as effective as possible.

The development modes of cloud-based clinical research databases

Origin of cloud computing

After years of development, database technology has gradually matured and is playing an important role in today's information society. However, it has also exposed some problems with its application, mainly in two aspects. First, there is an issue with updating data. Traditional databases lack real-time and initiated updates. Because there are mostly historical data in traditional data warehouses and the data retrieving cycle generally lasts several days or even a week, it is difficult to process data in real time. Meanwhile, traditional data warehouses use Extraction Transformation Loading (ETL) and adopt periodic batch updates in which the time and data of the update are preset and unable to change as needed. Second, the application scope and field of traditional data warehouses are rather narrow (10).

In recent years, the application of cloud computing in databases has resulted in new advances such as data clouds, real-time data, intelligent analysis of data, and others, providing new ideas for solving the above problems. So-called cloud computing refers to network technology that uses a computer as a carrier and the network and the Internet as bases to provide real-time services to users. This is an ad-hoc service, i.e., it provides what the user needs and charges service fees based on the user's usage, which greatly improves the rate of resource sharing (11). The essence of a cloud computing-based database is a large-scale centralized joint database management system, distributed across physical structures but logically belonging to the same system. The combination of cloud computing and databases has two forms: databases running in the cloud and cloud databases (12). The development of these two forms of databases is still in its infancy; while evolving in their own ways, they also bring more possibilities to database applications.

Data cloud

Data cloud technology is based on traditional storage and storage over the Internet, supplemented with the concept of unlimited resources, aiming to become another source of RAM and hard drive space for the global network of users that are connected at high speeds and synchronically share massive amounts of data via the Internet. Studies on the data cloud are still rather preliminary, and its applications are limited by many factors, such as cost, heterogeneity, security issues, the test of time, and others. Nevertheless, Internet storage, the prototype data cloud, has long been familiar to the public and is in wide use in many countries, presenting users with great convenience; examples include well-known internet hard drives, Chinese mobile communication, Mofile, and online storage (Dostor.Com, MediaFire, OmniDrive, etc.) (13). With the continuous development and maturation of data cloud technology, it will eventually "ease the burden" for PCs and "lose weight" for the Internet, which will make RAM and hard drive capacity expand without limits, thus bringing in an unlimited amount of shared resources, prompting the whole world to progress toward greater network connectivity.

Real-time databases

Another major advance in current database applications is the establishment and application of real-time databases. Real-time databases are still essentially data warehouses, but they are different from traditional databases in the most prominent feature of being real-time, reflecting the real-time changes in data warehouses: as long as there is a new event completing and generating data, the real-time database is able to capture these new data and update themselves, and the new data are immediately available. Different from the "snapshot" feature of the traditional data warehouse, a real-time data warehouse can synchronically reflect data changes in the business system (OLTP), thus making relevant analyses and decisions in real time. In short, real-time data warehouses effectively overcome the shortcomings of the traditional data warehouse, e.g., poor performance in real time and lack of capacity to provide flexible and timely tactical decisions for businesses; therefore, real-time data warehouses have broad prospects for development (10).

Intelligent analysis

The increasingly large amount of data requires more demanding data mining techniques, which has prompted the emergence of technologies in intelligent analysis (14). Widely used intelligent analysis techniques can be represented by Bayesian networks (15-17), mainly used in clinical studies with an extraordinarily large scale. Compared with conventional approaches to analysis, intelligent analysis analyzes clinical data from a complete novel perspective and has broad prospects for application. With the continuous improvement in computational performance, complex high-dimensional model algorithms represented by deep learning have gradually emerged (18). Without the need for human intervention, deep learning can automatically extract features layer-by-layer from huge amounts of data and form high-dimensional data models to simulate complex data, which has surpassed the traditional algorithms in the fields of image and voice recognition and will play a crucial role in the future field of intelligent analysis (19).

Although cloud computing, real-time databases, and intelligent analysis are high-tech areas that are still in exploratory stages, the translation and application of these new technologies to clinical research and analysis will advance clinical research models into a new era of big data.

Acknowledgements

Funding: This clinical study was supported by Fundamental Research Funds of State Key Laboratory of Ophthalmology (Grant No. 2015QN01), Young Teacher Top-Support project of Sun Yat-sen University (Grant No. 2015ykzd11), the Cultivation Projects for Young Teaching Staff of Sun Yat-sen University (Grant No. 12ykpy61) from the Fundamental Research Funds for the Central Universities, the Pearl River Science and Technology New Star (Grant No. 2014J2200060), Project of Guangzhou City, the Guangdong Provincial Natural Science Foundation for Distinguished Young Scholars of China (Grant No. 2014A030306030) and Youth Science and Technology Innovation Talents Funds in Special Support Plan for High Level Talents in Guangdong Province (Grant No. 2014TQ01R573), Key Research Plan for National Natural Science Foundation of China in Cultivation Project (No.91546101).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Li Q, Gao R, Lu F. The general procedure of clinical research data management and its management status quo. Proceedings of 13th National Conference of Clinical Pharmacology in China, 2012:522-5.
2. Cao W, Yan J. Database Review. Technology Management Research 2006;26:235-7.
3. Zhang Z. Databases and their development. Information and Documentation Services 1996;17:36-40.
4. Xu Z. Several questions about database construction. Journal of The China Society For Scientific and Technical Information 1994;13:365-9.
5. Wang T, He H, Luo Y, et al. Studies on the construction and applications of biological databases. Biotech World 2015;9:178.
6. Bazelier MT, Eriksson I, de Vries F, et al. Data management and data analysis techniques in pharmacoepidemiological studies using a pre-planned multi-database approach: a systematic literature review. Pharmacoepidemiol Drug Saf 2015;24:897-905.
7. de Waure C, Poscia A, Virdis A, et al. Study population, questionnaire, data management and sample description. Ann Ist Super Sanita 2015;51:96-8.
8. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. JAMA 2014;312:1024-32.
9. Platts-Mills TF, Jones CW. Reanalyses of trial results. JAMA 2015;313:92-3.
10. Jiang Z, Huang X. Studies on real-time data warehouse technology. Computer Systems & Applications 2007;17:91-4.
11. Yao S. Analysis of the application of database technology based on cloud computing. Computer CD Software and Applications 2013;16:296-7.
12. Wang L, Xu Y. The application and realization of a cloud-based database management system in higher education. Agriculture Network Information 2011;26:58-60.
13. Kuang S, Zhou Q, Liu X, et al. Key issues in the development of data cloud technology. Computer Science 2009;36:282-4.
14. Li G, Luo H. Studies on intelligent data analysis technology under big data. Technology Information, 2013;11:11-2.
15. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 2015;47:284-90.
16. Hampson LV, Whitehead J, Eleftheriou D, et al. Bayesian methods for the design and interpretation of clinical trials in very rare diseases. Stat Med 2014;33:4186-201.
17. Carreras G, Gorini G. Time trends of Italian former smokers 1980-2009 and 2010-2030 projections using a Bayesian age period cohort model. Int J Environ Res Public Health 2013;11:1-12.
18. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504-7.
19. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature 2015;518:529-33.

Cite this article as: Long E, Huang B, Wang L, Lin X, Lin H. Construction of databases: advances and significance in clinical research. Eye Sci 2015;30(4):184-189. doi: 10.3978/j.issn.1000-4432.2015.11.13