doi: 10.3978/j.issn.2095-6959.2015.
View this article at: http://dx.doi.org/10.3978/j.issn.2095-6959.2015.

数据库建设在临床研究中的重要作用和最新进展

龙尔平[1]*，黄冰洁[1]*，林晓瑜[1] 综述 王黎明[2]，林浩添[1] 审校

(1. 中山大学中山眼科中心，眼科学国家重点实验室，广州 510060；2. 西安电子科技大学软件学院，西安 710126)

[摘 要] 数据库是一门数据管理自动化的新技术，是数据管理最有效的手段，在当代临床研究中应用广泛。本文首先介绍了数据库的定义、类型和建设关键因素等基本概念。随后系统阐述了数据库建立、数据录入、审核和管理的流程，并从其在临床研究中的应用方法着手，介绍了数据库在临床研究中的应用现状，阐明数据库对临床研究水平的重要影响。最后，本文通过对临床数据的二次分析和以数据云为代表的云计算应用的讨论分析，阐述数据库在临床研究中的最新进展。

[关键词] 数据库；临床研究；数据管理；二次分析；云计算

Construction of databases: advances and significance in clinical research

LONG Erping[1]*, HUANG Bingjie[1]*, LIN Xiaoyu[1], WANG Liming[2], LIN Haotian[1]

(1. The State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou 510060; 2. School of Software, Xidian University, Xi'an 710126, China)

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

* 作者贡献相同

收稿日期 (Date of reception)：2015–11–24

通信作者 (Corresponding author)：林浩添，Email: gddlht@aliyun.com

基金项目 (Foundation item)：眼科学国家重点实验室青年加速计划项目 (2015QN01)，教育部高校 (中山大学) 青年教师重点培育项目 (2015ykzd11)，教育部高校 (中山大学) 青年教师培养项目 (12ykpy61)，广州市珠江新星项目 (2014J2200060)，广东省自然科学基金杰出青年项目 (2014A030306030) 和广东省高层次人才特殊支持计划“科技创新青年拔尖人才”(2014TQ01573) 国家自然科学基金重大研究计划培育项目 (91546101)。This clinical study was supported by Fundamental Research Funds of State Key Laboratory of Ophthalmology (2015QN01), Young Teacher Top-Support project of Sun Yat-sen University (2015ykzd11), the Cultivation Projects for Young Teaching Staff of Sun Yat-sen University (12ykpy61) from the Fundamental Research Funds for the Central Universities, the Pearl River Science and Technology New Star (2014J2200060), Project of Guangzhou City, the Guangdong Provincial Natural Science Foundation for Distinguished Young Scholars of China (2014A030306030) and Youth Science and Technology Innovation Talents Funds in Special Support Plan for High Level Talents in Guangdong Province (2014TQ01R573), Key Research Plan for National Natural Science Foundation of China in Cultivation Project (91546101), P. R. China.

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

数据库技术，诞生于20世纪60年代，是为了一定目的，在计算机系统中与特定的结构组织、存储和应用相关联的数据集合，是目前数据管理最有效的手段。数据管理对临床研究至关重要，贯穿整个研究过程。随着数据库技术在临床研究工作中广泛应用，不仅使研究数据的收集和存储更加便利可靠，而且使数据分析高效化，大大提高了研究效率，并推动了临床研究的飞速发展。然而，随着现代科技的发展，数据库应用领域出现了二次分析和云计算等新技术，这些新技术在临床研究分析中的应用，不仅给临床研究工作带来了挑战和机遇，且也带来了更多的便利和可能性。全面把握数据库建设在临床研究中的重要作用和最新进展，已成为当前我国广大临床研究人员最重要日常工作之一。

1 临床研究数据管理和数据库的定义、类型和建设的关键因素

临床研究数据管理，包括确认临床实验数据收集的完整性和正确性，确保实验数据对统计分析的支持，以及最后对实验结果的阐述和解释，贯穿整个临床研究过程。传统上，临床研究数据管理已具有国际比较认可的一般流程[1]，我们将其总结为图1。目前，临床实验数据获取方式有书面数据获取(paper data capture，PDC)和电子数据获取(electronic data capture，EDC)两类，总的来说，EDC具有提高实验效率和数据质量、节约成本、缩短研究周期、减少脱落、改善用户体验等优势，EDC工具的发展及在临床研究领域中的应用极大地推动了临床数据管理的电子化进程。

数据库是数据管理的高级阶段，它与传统的数据管理相比有明显的差别，主要体现在两个方面：一是使数据独立于应用程序，从而实现了数据的集中管理，并通过数据共享和减少数据冗余提高了数据的效益；二是建立起不同数据库之间的联系，从而反映出现实世界中信息的联系，使数据库不再只是单纯的数据集合[2]。数据库可按其性质分为三大类：文献型数据库、数值型数据库和事实性数据库。每一类数据库又可细分。文献性数据库包括专业文献库、馆藏书目库、联合目录数据库、专利文献数据库、全文数据库等。数值型数据库包括科学数据库、工程技术数据库等。事实型数据库包括经济与商情数据库、管理数据库、公用服务数据库等[3]。三大类数据库都在临床研究中有广泛应用。

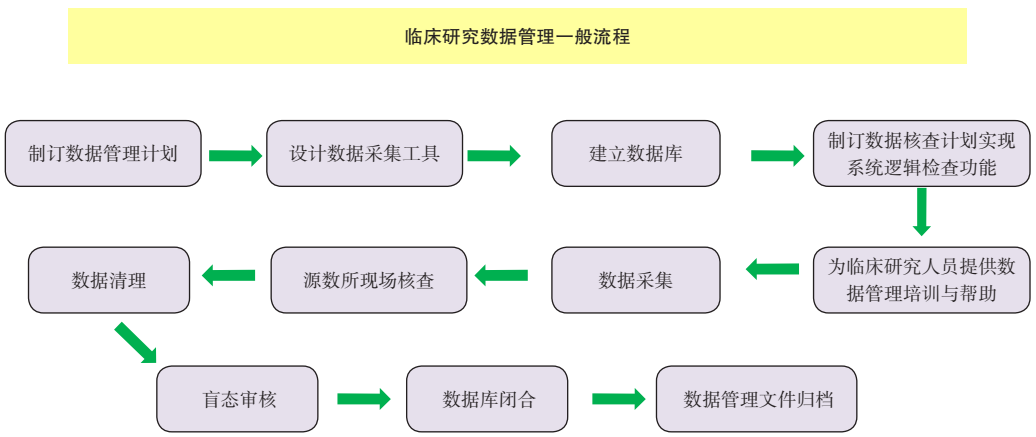

图1 临床研究数据管理一般流程图

Figure 1 General flowchart for clinical research data management

按照数据模型的进展，数据库技术可以相应地分为三个发展阶段：第一代的网状、层次数据库系统；第二代的关系数据库系统；第三代的以面向对象模型为主要特征的数据库系统。随着数据类型及数据关联复杂性的不断增加，单纯依靠根树结点搭建的网状及层状数据库已无法满足日益增长的数据库功能需求，现已基本淘汰。当今最流行是由关系式数据结构组成的第二代关系数据库，它通过将数据结构转化为二维关系形式，实现数据的分类、合并、连接和选取，并以此组合搭建具有复杂全面管理功能的数据平台。然而，不同应用领域对其数据库有特殊的功能需求，这是关系数据库所不能完全覆盖的。因此，拥有更加丰富数据模型及更个性化管理分析功能的第三代数据库开始崭露头角，但限于其高昂的成本，目前仅应用于军事等少数领域[2]。

数据库建设中有三大要素至关重要：管理、数据、技术。首先，管理是数据库建设的根基。它包括前期的决策和制订相应的政策、策略和措施，开发期的以软件工程为主体的软件开发管理和数据获取、质量控制；服务和维护期的数据安全与保密、更新、重组以及应用软件的三类维护管理(纠正性维护、适应性维护和完善性维护)。其次，数据是数据库建设的基础，数据源的可靠和稳定是一切的前提，因此必须对数据采集、录入、处理等各环节提出更高的要求。最后，技术是与数据库建设最直接的保障，其应用广泛而复杂，主要包括数据库技术、计算机技术和网络、通信技术[4]。

2 临床数据库的建立、数据的录入、审核和管理的流程

临床数据库的建立、数据的录入、审核和管理是非常繁复而缜密的过程。其基本流程如图2所示。然而，我们在具体实际过程，需要注意事项如下：1)确认数据的来源。主要包括研究地点和相关的实验室等。然后要确定需要收集的数据项目（研究变量）、定义变量(概念的定义，操作的定义)、数据库说明(数据类型，单/重复)、确认数据与原始资料一致的说明、数据确认说明(数值和数据范围的检查，缺失值，逻辑的检查)等。2)数据的收集与复核。数据收集基本要求是及时、完整、准确，详述如下：首先，“及时”是为了保证了原始数据的质量，也减少了以后复核的时间，“完整”就是要求收集所有研究对象的全部数据，“准确”首先要求设计的病例报告表应有较好的可操作性，应尽量采用计量化的指标，对软指标也应尽可能地做适当的量化；其次，保持实验室条件和操作人员的相对恒定；第三，要求临床医师或资料收集者在收集资料或填写病例报告表时需经过一定的培训；第四，要尽量减少从原始资料中过录数据，如果进行过录，应仔细查对，保证病例报告表与原始资料的一致。在数据收集完成后，还要进行数据复核，包括自我复查和监督、检查。3)编码工作。编码方法有人工编码、自动编码技术、字典：MedDRA，WHO-ART，ICD，COSTART。4)数据质量控制。常用的方法有数据跟踪、数据录入与确认、数据的核查、数据问题表和电脑程序检查等。5)数据录入前后的核查。临床数据的核查由有资格的医学核查员进行；检查和评估复杂的临床数据，发现数据中的细微差别；对数据管理员的评估进行质量控制检查。6)正式录入数据，建立数据库。录入工作要求是有问题的数据得到解决、数据录入已完成并得到确认或已抽样进行质量控制、所有有关安全性和有效性的主要数据都进行了质量控制核查、所有与研究方案不相符的事件都进行了报告和分类，并且澄清了相关的影响。符合要求的录入工作完成后可进行数据的整理并且传送至最终的数据库。7)设盲数据核查与数据库锁定。在确认了资料的真实性，盲态数据核查完成后，进行锁定，最后分析；与研究者一起进一步明确分析的目的和顺序；接受数据库并对资料进行初步的分析；进行初步数据分析；为最后的分析制定草案写成或修改统计分析计划书。数据库核查完毕应进行锁定，以防止误操作和未经授权的修改，并且准备一份临床研究报告所需要的数据列表。最后数据的传输和存档。将数据传输给申办者、统计人员、法规管理部门[5]。

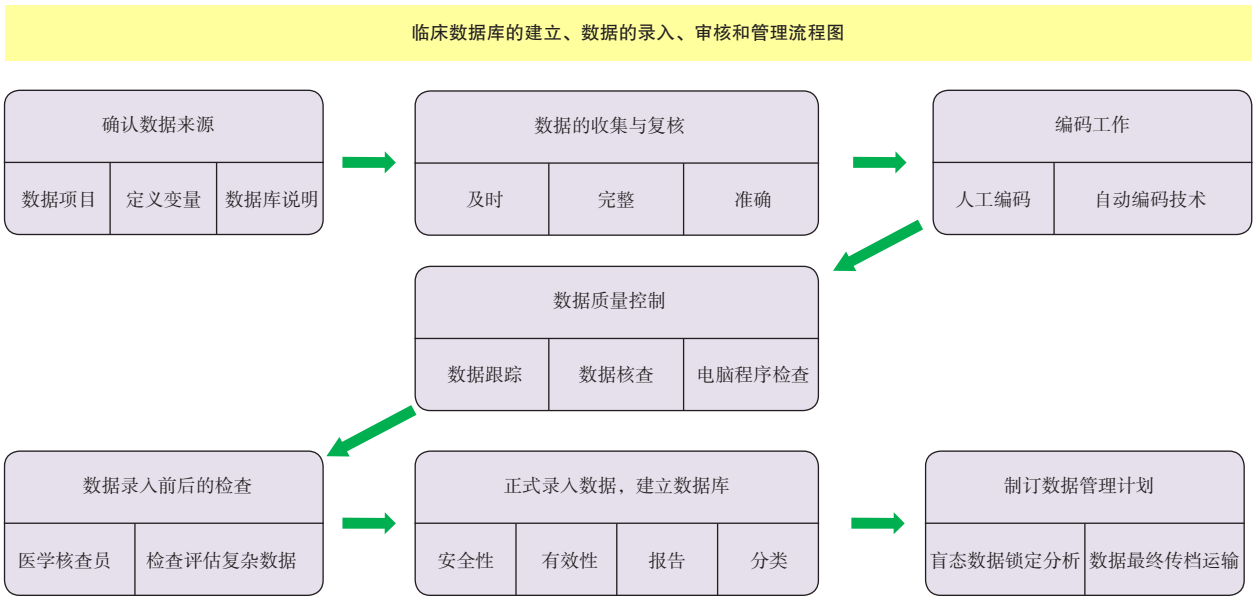

图2 临床数据库的建立、数据的录入、审核和管理流程图

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

3 临床数据库对临床研究水平的重要作用

临床数据库对临床研究水平的重要作用主要体现在两点：第一，临床数据库是数据分析和结论推导的前提；第二，临床数据库是衡量研究水平的依据之一。在临床研究中产生的庞大数据量，通过数据库的规整后可以大大提高研究效率以及数据的利用价值，并且在长期的研究进程中为研究者提供经验、依据。对数据库进行分析可以对临床研究进行检验，显示临床研究的意义以及其真实可靠性，有效地规范临床研究的设计和实施过程[6]。在临床研究中数据的收集和管理是实时进行的，任何数据管理的失误都可能切断原本紧紧相扣的临床实验的环节，造成不可祢补的损失。数据管理的质量反映出临床实验研究者对实验设计和实验计划的执行情况，也影射着临床研究者的科学态度和知识水平。数据管理也常常作为临床研究实施的主要检查内容，也是监视员和稽查人员的主要职责内容之一[7]。

4 基于临床数据的二次分析对研究结论的影响

二次分析是数据库对临床研究模式影响的最新进展之一。2014年，美国医学会杂志报道了一个具有里程碑意义的有关随机对照实验的二次分析的数据研究，研究者发现在37个案例中只有5个二次分析的作者与初始实验完全无关。在分析方法上，37个案例中总共出现了46种不同的方法——包括统计学方法和分析方法，且在二次分析最终得出的结果上，有35%的二次分析在患者的治疗方法上得出了与初始实验不同的结论[8]。

该实验代表性地说明：二次分析可对已有的临床研究数据进行重新分析，可能得出与初始实验相同或不同的结论。其作用在于检验和补充初始实验结果，或发现漏洞和错误。然而，这一分析方法由于发展时间较短，目前仍然存在以下几个隐患而引发了学术界的争议：1)二次分析过程可能对患者的隐私安全构成潜在的威胁，导致患者隐私暴露；2)二次分析中对数据集的不适当挖掘；3)二次分析研究中出现虚假伪造的分析结果；4)二次分析中商业机密的泄露[9]。

目前有文献[8-9]报道的二次分析研究并不多，分析实验的实施人、所采用的方法等都将影响分析结果。因此，目前二次分析的质量和可信度还令人担忧，出台更权威、影响更广泛的数据共享和二次分析的标准势在必行。只有制订统一、权威的标准，才能够减少二次分析研究中可能出现

的患者隐私被侵犯、商业机密泄露和实验用于攫取利益等各种问题，使二次分析研究发挥其最大的作用。

5 基于云计算的临床研究数据库的发展模式

5.1 云计算的产生

数据库经过多年的发展，其技术日趋成熟，在当今信息社会中发挥着重要作用。但在应用中也暴露出一些问题，主要体现在两方面：1)数据的更新问题。传统的数据库更新缺乏实时性和主动性。由于传统的数据仓库中大多是历史数据，数据抽取周期一般为数天甚至一周，因此很难对其进行实时性处理;而传统数据仓库采用ET'L周期性批量更新，更新的时间和数据都是既定的，无法根据需要进行更改。2)数据仓库的使用范围和应用领域狭窄[10]。

近年来，云计算在数据库中的应用带来了数据云、实时数据和数据的智能分析等新进展，为上述问题提供了解决思路。所谓云计算，指的是以计算机为载体，以网络、互联网为依托为用户提供实时服务的网络技术。这是一种有针对性的服务即用户需要什么就提供什么，并根据用户的使用量收取一定的费用，极大地提高了资源的共享率[11]。基于云计算数据库的实质是大规模集中式数据库管理系统的联合，在物理结构上是分布式的而在逻辑上属于同一系统。云与数据库的结合可分为两种：一种是运行在云中的数据库，另一种是云数据库[12]。此两种数据库的发展都处于起步阶段，在自身不断发展地同时，给数据库应用带来更多可能性。

5.2 数据云

数据云技术建立在传统存储和互联网存储的基础之上,并注入无限量概念,旨在成为全球网络用户的后台内存与硬盘,它将使每个人都能够通过互联网迅速链接和同步共享海量数据。虽然数据云的研究仍处在起步阶段，其应用依然受到成本、异构性、安全问题和时间考验等因素的限制，但数据云的雏形产品——互联网存储，却早已为大众所熟悉，并且已经在国内外发展得如火如荼，给用户带来巨大便利，如国内外知名网络硬盘中国活动通信Mofile、存储在线Dostor.Com、MediaFire、OmniDrive等[13]。随着数据云技术的不断发展与逐渐成熟,它终将为个人电脑“减负”和互联网“瘦身”,使它们的内存和硬盘无限扩充,从而为人们带来无限量的共享资源,并将促使全球迅速迈向更高阶段的网络化世界。

5.3 实时数据库

当前数据库应用的另一大进展是实时数据库的建立和应用。本质上实时数据库仍然是数据仓库，但它有别于传统数据库最大的特征是实时性，主要体现为数据仓库中数据的实时性变化：只要一有新的事件完成并产生数据，实时数据库就可以捕获这些新数据并完成更新，新数据立即可用。与传统数据仓库的“快照”形式不同，实时数据仓库能够同步反映业务系统(OLTP)中数据的变化，从而及时做出相关分析和决策。总之，实时数据仓库有效地克服了传统数据仓库实时差、难以为企业提供灵活及时的战术性决策等弊端，有着广阔的发展前景[10]。

5.4 智能分析

越来越庞大的数据量对数据挖掘分析技术也提出了更高的要求，智能分析的技术应运而生[14]。应用较为广泛的智能分析技术以贝叶斯网络为代表[15-17]，其主要应用于数量较大的临床研究。与传统的频率学派相比，以全新的角度解析临床数据，具有广泛的应用前景。随着计算机性能的不断提升，以深度学习为代表的复杂高维模型算法也逐渐崭露头角[18]。深度学习能从海量数据中自动逐层提取特征，不需人工干预，形成高维模型模拟复杂的数据，目前已经在图像和语音识别领域超越传统算法，必将在未来的智能分析领域扮演极为重要的角色[19]。

虽然，云计算、实时数据库和智能分析等计算均是一个正处在探索阶段的高科技领域，但是这些新技术在临床研究分析中的转化应用，将推动临床研究模式进入一个全新的大数据时代。

6 展望

基于循证医学的临床研究是目前医学诊疗的最关键部分。随着信息化和数据技术的不断提升，数据库的应用已成为临床研究质与量的重要保障基础。大样本长时序的队列研究，多维度

多源的生物医学大数据是未来临床研究的方向所在。因此，与大数据时代应运而生的计算云平台，智能分析技术将在经典统计学与分析方法基础上，开启一场颠覆传统医疗的精准革命。

参考文献

1. Li Q, Gao R, Lu F. The general procedure of clinical research data management and its management status quo[J]. Proceedings of 13th National Conference of Clinical Pharmacology in China, 2012: 522-525.
2. Cao W, Yan J. Database Review[J]. Technology Management Research, 2006, 26: 235-237.
3. Zhang Z. Databases and their development[J]. Information and Documentation Services, 1996, 17: 36-40. 2013X
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[J]. 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[J]. 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[J]. Ann Ist Super Sanita, 2015, 51: 96-98.
8. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data[J]. JAMA, 2014, 312: 1024-1032.
9. Platts-Mills TF, Jones CW. Reanalyses of trial results[J]. JAMA, 2015, 313: 92-93.
10. Jiang Z, Huang X. Studies on real-time data warehouse technology[J]. Computer Systems & Applications, 2007, 17: 91-94.
11. Yao S. Analysis of the application of database technology based on cloud computing[J]. Computer CD Software and Applications, 2013, 16: 296-297.
12. Wang L, Xu Y. The application and realization of a cloud-based database management system in higher education[J]. 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[J]. Computer Science, 2009, 36: 282-284.
14. Li G, Luo H. Studies on intelligent data analysis technology under big data[J]. Technology Information, 2013, 11: 11-12.
15. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts[J]. Nat Genet, 2015, 47: 284-290.
16. Hampson LV, Whitehead J, Eleftheriou D, et al. Bayesian methods for the design and interpretation of clinical trials in very rare diseases[J]. Stat Med, 2014, 33: 4186-4201.
17. Carreras G, Gorini G. Time trends of Italian former smokers 1980-2009 and 2010-2030 projections using a Bayesian age period cohort model[J]. Int J Environ Res Public Health, 2013, 11: 1-12.
18. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313: 504-507.
19. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.

本文引用： 龙尔平, 黄冰洁, 林晓瑜, 王黎明, 林浩添. 数据库建设在临床研究中的重要作用和最新进展[J]. 眼科学报, 2015. DOI: 10.3978/j.issn.1000-4432.2015.12.05

Cite this article as: LONG Erping, HUANG Bingjie, LIN Xiaoyu, WANG Liming, LIN Haotian. Construction of databases: advances and significance in clinical research[J]. Eye Sci, 2015. doi: 10.3978/j.issn.1000-4432.2015.12.05