多中心临床大数据平台建设及深度应用

朱立峰[1]，刘淑君[2]，陈德华[2]，乐嘉锦[2]
1. 上海交通大学医学院附属瑞金医院，上海 200025；
2. 东华大学计算机科学与技术学院，上海 201620

摘要
多中心临床研究是多中心、多学科对同一临床问题开展广泛协作临床研究的主要途径。传统多中心临床研究主要存在样本量偏小和临床科研相对封闭、开放程度不高的问题。为此，结合了新近兴起的大数据与云计算等技术，将物理上分散的各医院临床中心融合成逻辑上统一的临床大数据，构建了多中心临床大数据应用平台。介绍了多中心临床大数据平台的总体框架设计，详细阐述了平台各个子系统，分析了临床大数据平台的深度应用。

关键词
多中心临床研究；临床大数据分析与挖掘；临床科研随访系统

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018029

Construction and deep application of multi-center clinical big data platform

ZHU Lifeng[1], LIU Shujun[2], CHEN Dehua[2], LE Jiajin[2]
1. Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
2. School of Computer Science and Technology, Donghua University, Shanghai 201620, China

Abstract
Multi-center clinical research is the main approach for multi-center, multi-disciplinary, to develop some collaborative clinical researches on the same clinical issues. The traditional multi-center clinical research mainly has the disadvantages that small sample size and the clinical research is relatively closed and the degree of openness is not high. Therefore, the newly emerging technology, such as big data and cloud computing was combined to integrate clinical centers of physically dispersed hospitals into a logical and unified clinical data. On this basis, a multi-center clinical big data application platform was constructed. First, the overall framework of the multi-center clinical big data platform was designed, and then the subsystems of the platform were elaborated in detail. Finally, the deep application of clinical big data platform was introduced.

Key words
multi-center clinical research, analysis and mining of clinical big data, follow-up system of clinical scientific research

1 引言

近年来，多中心临床研究受到越来越多的关注。所谓多中心临床研究指的是由多个研究中心的临床医生或科研人员按照同样的研究设计、为同一个研究目的、协同完成的临床研究工作[1]。其中，研究中心可以是三级甲等医院，也可以是负责某个具体区域的社区医院。具体而言，在多中心临床研究中，临床科研由一个研究中心总体负责，担当牵头单位的角色，然后由多个研究中心的临床医生共同合作，按照同一个研究方案在不同的研究中心同时进行。这样，多位临床医生可不受地点的限制，在不同科室、不同医院按同一试验方案同时进行临床研究，协同完成各项研究工作。多中心临床研究实现了多中心、多学科对同一临床问题的广泛协作研究，对于发挥临床医生的学术优势、促进医学科学的发展具有重要的意义。

经过多年的努力，多中心临床研究已成为国内外各类医疗机构开展疾病临床研究的重要方法[2]。一方面，相对于单中心研究，多中心临床研究要求多个研究中心同时参与，可在较短的时间内遴选出临床科研所需的病例数；另一方面，相对于单中心研究，在多中心临床研究中多个中心入选的病例在病种病情分布等方面范围比较广。以糖尿病多中心研究为例，在确诊和治疗前期，患者多选择到三级甲等医院就诊，确定适合个体的治疗方案。治疗方案稳定后，患者大多会选择到社区卫生服务中心进行长期治疗和监督控制。因此，多中心的研究可以覆盖更多的糖尿病患者。

虽然多中心临床研究已被众多的医院、科研机构、临床医生广泛采纳，但是在实际应用中，多中心临床研究也具有许多不足之处。

（1）样本量往往偏小（即参与临床科研的病例数偏少）

在选取参与科研的样本病例上，当前的多中心临床研究通常采取一种协商遴选的办法，即各个研究中心的临床医生根据临床科研目标提供候选病例，再由总负责的研究中心确定目标样本。这种协商遴选的方法难以在较短的时间内收集足够多的病例[3]。当前的多中心临床研究的样本数量规模大致为几百，样本规模偏小，病例数不多。

（2）临床科研相对封闭、开放程度不高

在临床科研上，多中心临床研究需要科研团队之间共享科研数据和临床研究设计。当前的多中心临床研究在组建科研团队上大多采用如下3种方式：针对某一研究问题正式指派组成团队；有相同研究兴趣的同事组成团队；先决定团队的带头人，再由带头人发现团队成员。无论采取哪种方式组织团队，临床科研数据的搜集与共享都依赖于参与多中心临床研究的医院、科室、临床医生，这使得临床科研相对比较封闭。

因此，本文针对当前多中心临床研究存在的“样本量偏少、相对封闭”的缺点，结合新近兴起的大数据与云计算等技术，将物理上分散的各医院临床中心融合成逻辑上统一的临床大数据，在此基础上构建多中心临床大数据应用平台，并将之运用于多中心的临床研究实践中，实现多中心临床研究向“全样本、多学科、多病种、开放性”转变，探索基于多中心临床大数据的应用示范。

本文首先给出多中心临床大数据平台的总体架构，然后详细阐述临床随访系统、平台核心系统，最后阐述临床大数

据平台在内分泌和高血压等疾病上的深度应用。

2 多病种、多中心临床大数据平台总体框架

本文所提的多病种、多中心临床大数据平台的总体架构如**图1**所示。平台总体框架由数据基础层、临床随访系统、平台核心系统、应用系统构成。

- 数据基础层是整个平台的数据来源，多病种、多中心临床大数据主要来源于医院临床数据仓库、其他大型医院临床数据仓库及社区卫生服务中心临床数据仓库。
- 临床随访系统的目的是为专科医生进行临床科研随访提供支持，包括建立随访队列、形成随访表单、导出随访数据、对患者随访进行提醒跟踪。
- 平台核心系统是临床大数据平台的核心层，由3个子系统组成，即多中心临床大数据整合子系统、多中心临床大数据存储子系统、多中心临床大数据分析挖掘算法子系统。
- 应用系统为平台面向用户（包括医生和患者）提供的疾病分析应用服务，本文选择内分泌、高血压、肿瘤等若干病种作为研究对象，通过对多中心临床大数据进行分析与挖掘，更好地为医生和患者提供辅助诊疗服务。

3 临床随访系统

临床随访系统方便专科医生筛选随访队列，帮助专科医生自助设计随访表单，并在完成随访表单填写时，高效、准确地获取患者临床数据，如就诊记录、医嘱、检验检查报告、病理报告、手术麻醉记录、病史记录等，并提供随访表单数据手工录入功能，帮助专科医生获取和浏览患者在院内的全部诊疗记录，为科研与临床诊疗结合提供流程优化支持[4]。具体而言，临床随访系统包括以下4项。

- 随访信息维护。通过输入随访名称、课题名称、课题组成员、课题起始时间等信息创建随访队列，并基于随访队列维护课题组成员，可以添加账户或冻结账户，并通过结题管理维护队列状态。
- 随访表单管理，编辑管理表单模板名称、用途、适用的随访队列。每个随访队列可对应多个表单模板，主要可分为首次随访记录、住院随访记录、门诊随访记录等。
- 随访提示为临床医生提供如下功能：对于随访队列中的患者，如果超过预定时间（如门诊、住院一个月）无就诊记录，在随访队列中的负责医生登录时进行提示；如果存在就诊事件（门诊、急诊、住院），但尚未填写对应的随访表单，在随访队列中的负责医生登录时进行提示。
- 随访信息库提供网页形式，可由院外科研人员进入某一随访队列，添加患者信息，录入随访表单或者批量导入随访数据。

4 平台核心系统

本节阐述的多中心临床大数据平台的核心系统包括多中心临床大数据整合子系统、多中心临床大数据存储子系统和多中心临床大数据分析挖掘算法子系统。

4.1 多中心临床大数据整合子系统

临床数据整合技术旨在将现有的

不同临床信息系统产生的临床数据汇总到统一的信息平台，实现患者标识统一以及患者临床信息的统一浏览发布。目前，笔者所在单位拥有医院信息系统（HIS）、实验室信息系统（LIS）、放射科信息管理系统（RIS）、电子病历（EMR）、病理、手麻、心电等业务系统，数据库管理系统采用了MS SQL Server和 Sybase ASE，为了将数据抽取、加载和转换（extraction-loading-transformation，ETL）过程对实时业务系统的影响降到最低，采用规范化数据存储（normalized data store，NDS）+多维数据存储（dimensional data store，DDS）的数据流架构，临床数据整合流程如**图2**所示。ETL是构建数据仓库的重要一环，用户从数据源抽取所需的数据，经过数据清洗，最终按照预先定义的数据仓库模型，将数据加载到数据仓库中。

其中，Stage为源系统数据的一个副本，从Source到Stage的ETL过程并不对数据进行删减、修改或整合，仅负责抽取和加载，而从Stage到NDS的ETL过程则需要按照数据仓库建模的需求对数据进行一定的清洗和转换[5]。

4.2 多中心临床大数据存储子系统

多中心临床大数据存储子系统是整个平台的存储基础，为数据筛选、过滤、处理、分析提供原始数据，要求容量大、效率高、安全性强。云存储是通过集群应用、网络技术分布式文件系统等将不同种类的存储设备协调工作，共同对外提供数据存储和访问功能的一个系统，它以数据存储和数据管理为核心[6]。本文采用云计算的数据管理平台实现对多中心临床大数据的多元化存储与访问，

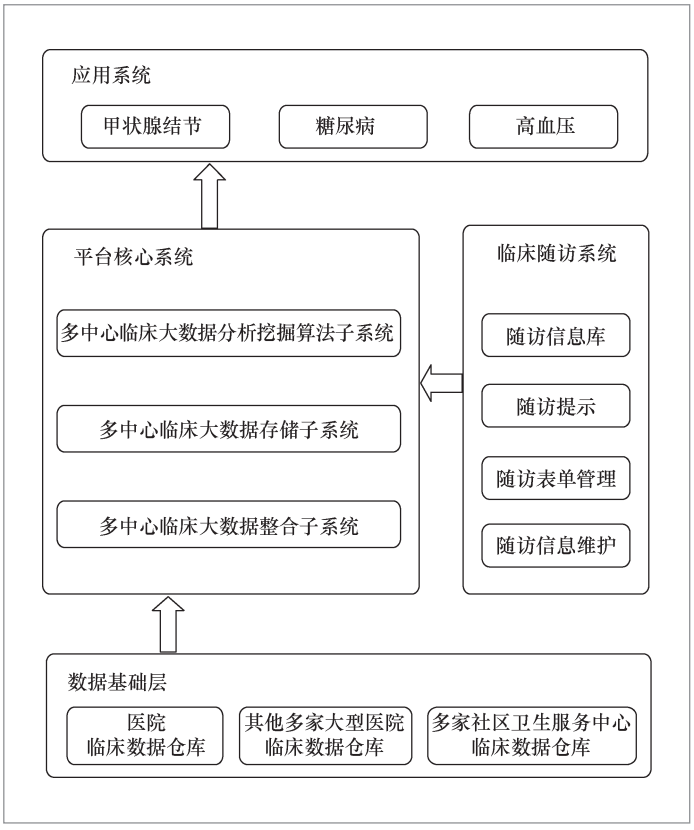

图1 多病种、多中心临床大数据平台总体架构

提供线性扩展的分布式存储能力。多中心临床大数据存储系统通过整合关系数据库、NoSQL数据库、文件系统等异构存储模式，实现多中心异构临床数据按需、按数据特性存储，支持大容量存储空间，存储效率快，可扩展性强，并通过统一的数据总线实现跨存储设施的统一访问。此外，借助并行数据处理技术对吞吐量、并发量以及可用性进行增强。

多中心临床数据种类很多，针对不同的临床科研需求和数据特点，遵循HL7标准对各种来源的多中心临床数据进行语义转换、结构变换，然后调用统一数据访问接口将其存储到合适的底层存储系统中。多中心临床大数据存储子系统考虑到异构数据的特点，采用了不同的底层存储设施，具体包括MongoDB、MySQL、

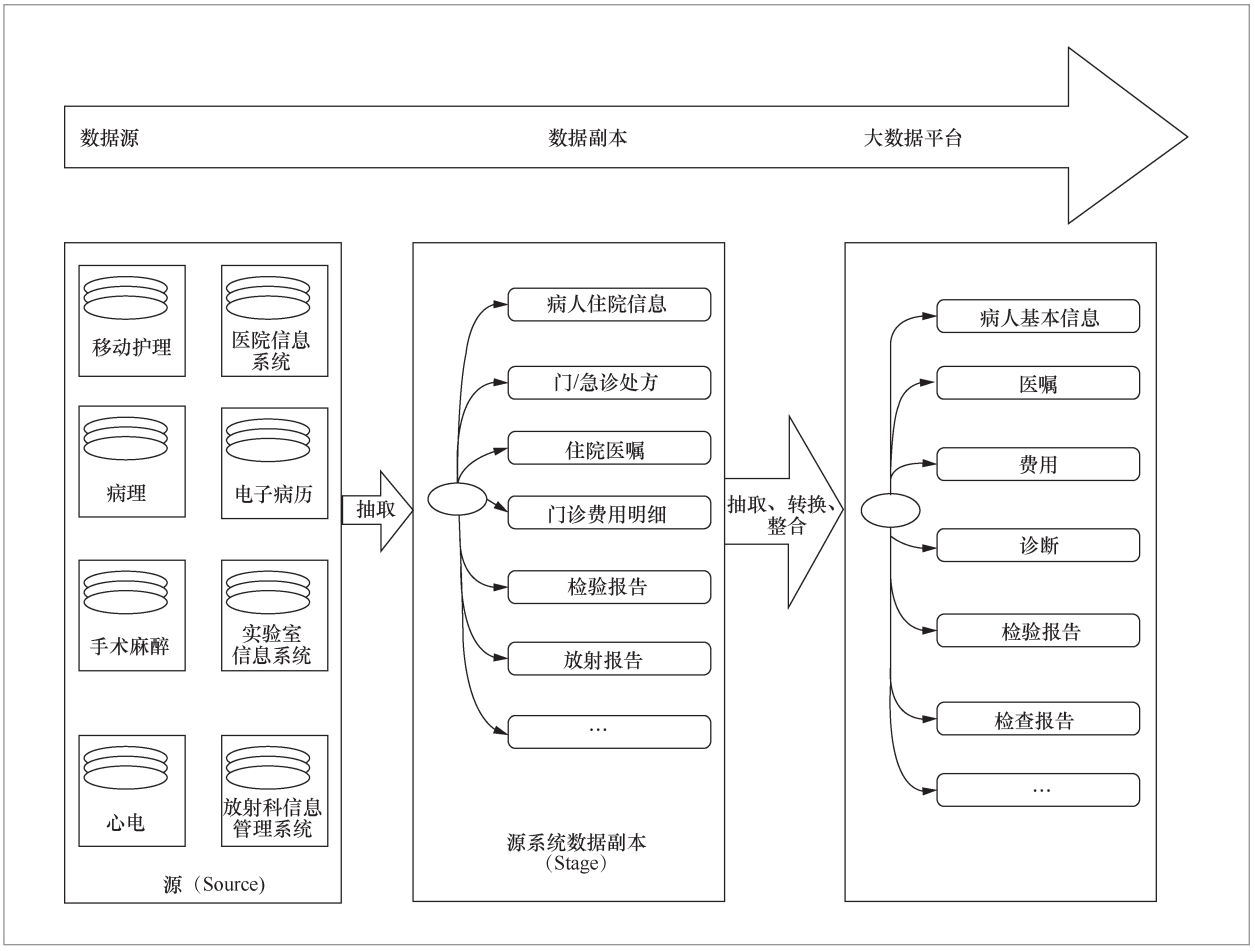

图 2 数据整合流程

HBase、Hadoop分布式文件系统（HDFS）等关系数据库、列存储数据库和分布式文件系统。其中，MongoDB、HBase主要用来存储半结构化数据，如检验/检查报告、出院小结等；MySQL主要用来存储结构化数据，如病人信息、处方等；HDFS主要用来存储非结构化的二进制文件，如来自影像归档和通信系统（PACS）的放射检查胶片图像。

4.3 多中心临床大数据分析挖掘算法子系统

多中心临床大数据分析挖掘算法子系统为疾病临床数据分析挖掘系统提供多维分析、分类、聚类、预测、回归分析和深度学习算法支持。在该系统中，共划分为多维分析算法、数据挖掘算法、深度学习预测算法三大基础模块。

（1）多维分析算法

多维数据分析首先需要建立多维数据集，由于其具有很多维度的特性，多维数据集通常被形象地称为数据立方体（cube）。多维数据集是一个数据集合，通常先从数据仓库中选取若干数据子集，再组织和汇总成由多个维度和度量值定义的多维结构，多维数据分析可对以多维形式组织起来的数据立方体进行上卷、下钻、切

片、切块、旋转等多种分析操作，以便剖析数据，使分析者、决策者能从多个角度、多个侧面观察数据库中的数据，从而深入了解包含在数据中的信息和内涵[7]。

（2）数据挖掘算法

数据挖掘算法模块提供了包括分类、聚类、关联规则、回归分析等机器学习算法的统一注册、应用与注销管理，用于针对特定数据集的挖掘分析，实现临床的深度分析、预警与预测。

数据挖掘算法模块中集成了随机森林、支持向量机、神经网络、决策树等分类算法和*K*-means聚类算法、逻辑回归、线性回归和关联分析等算法[8]。

（3）深度学习预测算法

深度学习预测算法模块集成了卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）和长短期记忆（long short term memory，LSTM）网络[9]等算法。

5 多中心临床大数据平台深度应用

在多中心临床大数据平台的基础上，进一步结合内分泌学科、高血压等学科的优势，开展临床大数据的深度应用。

（1）面向甲状腺肿瘤的临床医疗大数据的知识挖掘系统

临床中患者肿瘤的良恶性诊断结果的准确性主要依赖于医生的主观判断，不同医生的临床经验和医院医疗水平等存在差异，势必会造成不同程度的误诊。为了提高医生临床诊断的准确性，简化不必要的检查过程，降低误诊率，通过大数据分析方法实现医疗智能化已成为研究的热点。本文设计开发的系统，通过对甲状腺肿瘤治疗过程中的临床数据进行挖掘分析，实现了对甲状腺肿瘤良恶性诊断的准确预测。

（2）糖尿病转归及风险评估系统

糖尿病作为一种常见的慢性病，有治疗周期长、并发症多、病情反复等特点。当前的糖尿病的诊断还是以传统的诊断方式为主，每个门诊患者和医生接触的时间有限，医生无法在每个病人身上花费大量时间深入探究患者的个体情况和诊疗背景，这导致医院和医生无法更好地在后期监测患者状况。在有关糖尿病预测的实验中，可以将样本数据根据要求按照一定比例划分成训练集、测试集和独立样本集，进行糖尿病的预测和建模，对糖尿病高危人群提供有效、准确的帮助[10]。本文针对患者糖尿病生化指标预测这一问题，构建了一个基于缓存神经网络的糖尿病生化指标变化趋势预测模型。该模型特点是在隐藏层中增加计算结果缓存区域，以应对数据模型的不规则时序特性和内部复杂的相互关系，从而达到辅助诊疗的目的。

（3）基于数据挖掘的高血压患病趋势分析及控制

依托临床大数据平台，对高血压科近15年的住院患者的约500万条临床数据及门诊患者的约700万条临床数据进行收集整理，并进行多维度分析。找出该病种的患病趋势，并通过对降压药处方的分析，指导临床更加有效地防治高血压。通过临床数据分析，识别高血压患者靶器官（心、脑、肾、血管等）损伤，获取高血压患者的心血管及其他指标参数，基于这些参数，进行数据评估患者心血管风险，降低心脑血管病发生率[11]。

6 结束语

本文针对传统多中心临床研究存在的

问题，逐步推进多中心临床大数据平台建设和深度应用工作，完成了多中心临床大数据平台建设，并对多中心临床大数据平台中临床随访系统、核心系统和疾病分析应用系统等进行技术验证，建立了可行的技术框架。

参考文献:

[1] 丁洁. 关于多中心临床研究的思考[J]. 中华儿科杂志, 2011(11): 801-803.
DING J. Thoughts on multi-center clinical research [J]. Chinese Journal of Pediatrics, 2011(11): 801-803.

[2] 李娌, 魏玉萍, 钱芳桥, 等. 加强多中心临床研究的组织管理水平促进多中心临床研究质量提高[C]//2011第三届全国医学科研管理论坛暨江苏省医学科研管理学术年会论文集, 2011年10月1日, 南京, 中国. [出版地不详: 出版者不详], 2011: 207-210.
LI L, WEI Y P, QIAN F Q, et al. strengthen the organization and management of multi-center clinical research to promote the quality of multi center clinical research to improve the [C]//2011 3rd National Medical Scientific Research Management Forum and the Jiangsu Provincial Medical Scientific Research Management Academic Annual Meeting, October 1, 2011, Nanjing, China. [S.l.:s.n.], 2011: 207-210.

[3] 万霞, 李赞华, 刘建平. 临床研究中的样本量估算:(1)临床试验[J]. 中医杂志, 2007(6): 504-507.
WAN X, LI Z H, LIU J P. Sample size estimation in clinical research: (1) clinical trials[J]. Journal of Chinese medicine, 2007(6): 504-507.

[4] 崔洁, 朱立峰, 左铭. 临床科研随访系统需求与策略[J]. 中国数字医学, 2016(12): 36-37, 71.
CUI J, ZHU L F, ZUO M. Requirements and strategies for clinical research follow-up system[J]. Chinese Digital Medicine, 2016(12): 36-37, 71.

[5] 苌程. 基于ETL的数据集成过程模型[D]. 杭州: 浙江大学, 2010.
CHENG C. Data integration process model based on ETL[D]. Hangzhou: Zhejiang University, 2010.

[6] 陈杰. 大数据场景下的云存储技术与应用[J]. 中兴通讯技术, 2012, 18(6): 47-51.
CHEN J. Cloud storage technology and application in big data scenario[J]. ZTE Technology, 2012, 18(6): 47-51.

[7] ALPAY D. System theory, the Schur algorithm and multidimensional analysis[J]. Operator Theory Advances & Applications, 2007, 21(3): 133-150.

[8] DAN J, QIU J, XIANG G, et al. A synthesized data mining algorithm based on clustering and decision tree[C]// International Conference on Computer and Information Technology, June 29-July 1, 2010, Bradford, UK. Piscataway: IEEE Press, 2010: 2722-2728.

[9] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.

[10] 洪烨. 基于机器学习算法的糖尿病预测模型研究[D]. 哈尔滨: 哈尔滨工业大学, 2016.
HONG Y. Research on diabetes prediction model based on machine learning algorithm[D]. Harbin: Harbin Institute of Technology, 2016.

[11] 苏杰. 基于数据挖掘技术的高血压疾病评估方法研究[D]. 杭州: 浙江大学, 2006.
SU J. Research on hypertension disease assessment method based on data mining technology[D]. Hangzhou: Zhejiang University, 2006.

作者简介

朱立峰(1976-),男,上海交通大学医学院附属瑞金医院高级工程师,主要研究方向为医疗信息管理和医疗大数据。

刘淑君(1995-),女,东华大学计算机科学与技术学院硕士生,主要研究方向为大数据和文本挖掘。

陈德华(1976-),男,博士,东华大学计算机科学与技术学院副教授,主要研究方向为大数据和智慧医疗。

乐嘉锦(1951-),男,东华大学计算机科学与技术学院教授、博士生导师,主要研究方向为数据库与数据仓库、大数据与智慧医疗。

收稿日期: 2018-03-02
基金项目: 上海市科技发展基金资助项目(No.16JC1400802, No.16JC1400803)
Founndation Items: Shanghai Science and Technology Development Fund (No.16JC1400802, No.16JC1400803)