呼吸专科数据平台多中心数据集成与应用*

方 莹 陈 智

（1 广州医科大学附属第一医院 广州 510120
2 国家呼吸系统疾病临床医学研究中心 广州 510120
3 暨南大学第一临床医学院 广州 510632）

（1 广州医科大学附属第一医院 广州 510120
2 国家呼吸系统疾病临床医学研究中心
广州 510120）

简文华 张冬莹 郑劲平

（广州医科大学附属第一医院 广州 510120）

〔摘要〕 构建呼吸系统疾病大数据应用平台，建立呼吸系统疾病标准数据集规范，以云数据平台整合多中心呼吸专病数据，介绍基于该平台实现的数据采集、处理、交互和应用。

〔关键词〕 医疗大数据；数据共享；数据元；数据应用
〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2021.02.011

Multi－center Data Integration and Application of Respiratory Data Platform

FANG Ying, 1The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, 2National Clinical Research Center for Respiratory Disease, Guangzhou 510120, 3The First Clinical Medical College, Jinan University, Guangzhou 510632, China; CHEN Zhi, 1The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, 2National Clinical Research Center for Respiratory Disease, Guangzhou 510120, China; JIAN Wenhua, ZHANG Dongying, ZHENG Jinping, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China

〔**Abstract**〕 The big data application platform for respiratory diseases is built, the standard data set specification for respiratory diseases is established, the multi－center data for respiratory diseases is integrated with the cloud data platform, and the data collection, processing, interaction and application based on this platform is introduced.

〔**Keywords**〕 medical big data; data sharing; data element; data application

1 引言

1.1 研究背景

随着信息技术进步和区域化医疗发展，多中心、跨区域、大样本的“医疗＋大数据”复合学科发展模式成为近年临床研究热点，在这一过程中建立了一系列信息化管理系统，提高医院临床、科研各部门数据管理和办公效率。由于各部门业务重点、信息化需求不一，从而产生各类型信息化系统，数据类型、系统框架建设时间和开发商不同导

〔收稿日期〕 2020－07－06
〔作者简介〕 方莹，硕士，发表论文 1 篇，参编论著 1 部，参与专利发明 3 项；通讯作者：郑劲平，教授，主任医师，博士生导师。
〔基金项目〕 国家重点研发专项“呼吸系统疾病临床研究大数据与生物样本库平台”（项目编号：2018YFC1311900）。

致数据接口不规范、整合难度大，信息系统间孤立存在，业务数据难以共享，增加数据对接操作复杂程度，引发各类运维管理问题，产生信息孤岛现象。针对多中心数据规范化管理，提出基于呼吸系统疾病数据标准集建立的大数据应用云平台，整合多源异构数据，提供数据传输、交互服务及应用功能模块。

1.2 医疗大数据在临床研究中的应用

呼吸系统慢性疾病发病率、死亡率呈上升趋势，其中慢性阻塞性肺疾病已成为我国慢病死亡率第 4 大疾病，给社会带来经济和医疗负担[1]。我国呼吸系统疾病患者人数众多，临床数据资源丰富，近年来大型多中心登记注册的临床研究项目逐渐增加，依靠大量具有地域代表性的医疗大数据，对发现真实世界研究（Real World Study，RWS）中呼吸系统疾病患者临床特征、挖掘疾病早期发病规律、建立规范化临床诊疗流程及随访康复管理具有重要作用。医疗大数据的发展为真实世界登记注册研究和临床大样本随机对照试验（Randomized Controlled Trail，RCT）提供更丰富的多源数据，保证 RWS 数据多元性、重要性和时效性[2]。

2 呼吸系统疾病大数据应用平台概况及建设目标

2.1 概况

广州医科大学附属第一医院、国家呼吸系统疾病临床医学研究中心自 2018 年初部署呼吸系统疾病大数据应用平台，依托国家临床研究中心已辐射 18 家分中心单位数据接入，形成全国多中心、具有地域代表性的呼吸系统疾病大数据网络平台，数据存储量达到 230 万人次，涵盖呼吸系统疾病电子病历、胸部影像学图像（CT 及 X 光）、肺功能报告、呼吸系统疾病生物样本数据及单病种登记注册研究数据库。

2.2 建设目标

综合医疗信息化建设现状和大数据在临床研究中的价值，呼吸系统疾病大数据应用平台以打破信息孤岛、实现多中心数据互联互通为目标，建设标准化、规范化、临床医学与生物信息结合的统一数据交换与应用平台。

3 建设内容

3.1 总体架构（图 1）

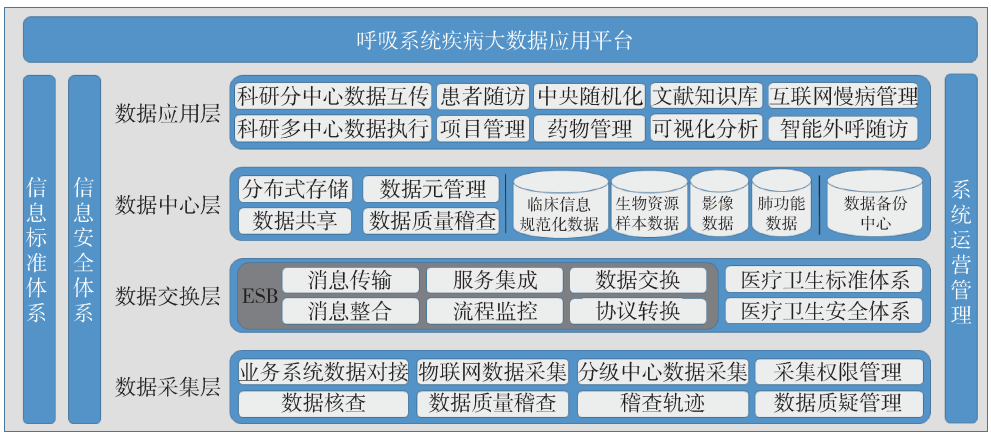

图 1 呼吸系统疾病大数据应用平台总体框架

平台搭建采用云端部署方式，使用第 3 方运营商“医疗云”专用云端服务环境，Hadoop 分布式计算框架，为海量数据提供存储和计算服务。根据国内外通用指南、专家共识建立呼吸专科标准数据元体系，提供统一数据标准和接口标准，实现不同业务系统、网络成员单位数据与呼吸系统疾病大数

据应用平台的有效集成与信息共享，通过医疗云平台充分保证接入系统的安全性，以满足平台数据交互及安全需求。平台根据数据采集、交互、应用处理流程，由数据采集层、数据交换层、数据中心层及数据应用层 4 个核心层面组成，数据应用模块包括呼吸专科数据采集及后结构化处理、数据元管理、文献知识库、呼吸专科数据仓库、数据分发与集成应用和数据安全审计管理，建立在基于数据延伸的平台应用，如呼吸慢病登记随访管理、科研项目管理、数据可视化分析和临床辅助决策支持系统。。

3.2 基于标准数据集的数据采集处理流程

3.2.1 数据采集 完成多中心数据采集、清洗、脱敏、映射、结构化处理等规范化工作，对电子病历中文本内容和胸部影像学图形数据进行提取和指标化处理。包括数据采集、同义词归一、数据元管理、质量稽查、可视化处理等一系列流程管理。应用提取 - 转换 - 加载（Extract - Transform - Load，ETL）过程管理，即数据采集、转换及加载，作为构建数据仓库的基础工具。对各类呼吸系统疾病临床诊疗数据，建立基于临床工作站业务信息系统的自动采集程序，形成单病种专科数据仓库，向呼吸系统疾病大数据应用平台各类应用提供数据支撑服务，建设临床信息规范化数据中心、生物资源样本数据中心、呼吸影像数据中心、科研随访管理等各类数据应用中心。

3.2.2 数据处理 保障数据应用质量的核心环节。针对多中心、大样本临床电子病历数据，对采集的字段进行顶层设计，参照临床数据交换协会（Clinical Data Interchange Standards Consortium，CDISC）标准、《中国公共卫生信息分类和基本数据集》、国际疾病分类第 10 次修订版本（International Classification of Diseases，ICD - 10）、SNOMED CT、LOINC 等国内外信息处理标准，结合相关呼吸系统疾病术语规范、国内外诊疗指南、专家共识等多种医学标准，联合国家呼吸系统疾病临床医学研究中心成员单位建立呼吸系统疾病公共数据元和专科病种数据元[3]，明确呼吸系统疾病数据元的中英文名称、定义、变量类型、值域、来源以及数据元间的树状结构。数据元的类型涵盖：人口统计学信息；疾病相关症状、体格检查、临床诊断、危险因素史、病历信息；检查检验相关肺功能检查、影像学检查、实验室检查；治疗相关吸入呼吸用药、口服呼吸用药和其他干预措施；病情评估相关临床评估量表、卫生费用和临床依从性；临床研究相关随访预后情况、不良事件和生物样本信息。在文本数据处理方面，应用自然语言处理技术（Natural Language Processing，NLP）和呼吸系统疾病标准数据元，采用医学标注和深度学习对采集数据进行后结构化处理，提取病历文本、检查检验报告中呼吸系统疾病关注的字段，如对住院病历中现病史症状、症状时间进行标注。形成临床信息规范化数据仓库，为中心数据共享提供标准化支持。在图形数据处理方面，利用图形处理器（Graphics Processing Unit，GPU）集群建立图形计算功能，提供 2D/3D 可视化工具，对胸部影像图像中的病灶进行标注，提取肺气肿、结节、支气管扩张等影像学数据信息，进行深度神经网络模型训练，提供低假阳性和高可靠性的检测算法。

3.3 数据交换

3.3.1 患者主索引（Enterprise Master Patient Index，EMPI） 基于采集的数据，使用对核实患者身份信息具有唯一标识功能的姓名、性别、身份证号、ID 号、医疗卡号、手机号码等字段作为核心参数，住址、职业、工作地址等字段作为参考参数，附带就诊医院信息，建立 EMPI 上传至云平台，方便各个应用间信息调取和整合。

3.3.2 企业服务总线（Enterprise Service Bus，ESB） 在传统信息系统数据交互技术中存在信息系统之间数据接口无法复用、接口专用性高等难题。ESB 提供标准形式的数据开放交互服务，通过中间件实现数据业务梳理及接口规范制定，数据交换打破院内与院外、各业务系统之间障碍，使得数据通信更加原子化和公式化[4]。ESB 使用 WebSphere MQ 消息队列（Message Queue，MQ）为数据负载工具，采用分布式部署将各个应用系统连接至

ESB，数据服务方应用系统通过 ESB 传输信息到达数据需求方业务系统，数据需求方从对应的 WebSphere MQ 队列获取数据消息，实现云平台上应用系统间互联互通。

3.4 安全防护

3.4.1 概述 云平台汇集各地区大量脱敏后的临床资料，信息安全尤为重要。随着《国家安全法》和《密码法》的颁布，需要充分重视建立网络信息安全体系。在数据保护方面应用国密算法 SM4 和 MD5 进行加密保护，在数据传输方面借助第 3 方通讯服务商的医疗云服务器平台及电路专线达到数据传输专网专用，一对一连线。

3.4.2 数据脱敏 应用国密算法 SM4。SM4 是我国无线局域网标准 WAPI 所采用的分组密码法算法，密钥长度和分组长度均为 128 位，用于通信加密、数据加密等[5]。对数据进行 SM4 算法加密处理后进行 MD5 算法的不可逆加密，不直接生成具有唯一性的脱敏数据，达到过程中加密效果。

3.4.3 网络安全 基于多业务传送平台（Multi-Service Transfer Platform，MSTP），通过以太网接口为平台提供点对点数据专线连接。使用专有线路数据传输环境，保障数据传输稳定性和安全性，加快信息数据传输速度[6]。

4 应用成效

4.1 单点登录门户

基于 CAS 软件技术建立统一身份认证门户，云平台用户只需使用 1 套账户密码可访问基于平台搭建的各个应用系统。登入统一身份认证门户时，账户与 CAS Server 以全局会话作为身份认证依据，应用系统与 CAS Server 以服务票据（Service Ticket，ST）作为验证依据，账户和应用系统之间以 Cookies 建立连接，用户无需反复登录操作即可直接访问，实现统一身份认证功能。

4.2 多中心数据共享，建立呼吸系统疾病标准数据元

实现多中心数据汇总，提供可视化实时数据展示大屏，可实时查看各医疗机构基本数据，如就诊量、病情评估随访管理、项目开展情况，加强医院间信息互换，促进多中心科研项目合作。同时基于呼吸系统疾病开展专病数据元指标库标注工作，随着更多网络单位的加入和系统应用的拓展，数据颗粒度将变得更加精细，协同分中心成员进行数据标准化标注，开放共享数据资源。

4.3 支撑多中心临床研究项目开展和规范化登记注册管理

建立多中心规范化呼吸慢病管理登记随访流程，应用智能语音外呼、公众号、APP 和短信提醒功能，定时向管理的呼吸慢病患者推送定期返院随访日程、用药提醒等信息，借助智能化随访管理平台提高患者随访依从性和疾病知晓情况。

4.4 协助提升呼吸病学科诊治水平

形成多中心报告质控、远程会诊、手术示教人工智能诊断及知识库共享等应用，推进呼吸专科检查技术如呼吸介入、胸部影像学、呼吸睡眠监测和肺功能检查在全国各级医院的普及，发挥传、帮、带作用。建立呼吸专科检查结构化模板，建立 3 级报告审核机制，推动优质资源向基层医院下沉，提升呼吸系统疾病诊治能力[7]。

5 结语

近年来医疗大数据快速发展，具有很大挖掘潜力，但同时信息孤岛、标准化不一等问题未能解决。崔春舜[8]等对国际健康大数据研究计划发展及启示的研究中指出，国内仍未形成统一标准及数据共享机制，需要多部门、多学科合作，加快医疗健康大数据应用。呼吸系统疾病大数据应用平台针对不同医疗机构及数据使用者，对海量医疗数据进行高效管理、挖掘和分析。在此基础上平台应面向公众、社会各类机构提供云服务环境，为医疗卫生事业、健康产业深入发展提供有力支撑。

（下转第 83 页）

已建立多个医学资源数据库[10]。智慧医院图书馆建设应重视数据库检索培训，开展中外文医学数据库资源整合、文献查重和查引、文献远程传递、课题查新跟踪、专题情报、文献翻译、为患者推送各类健康宣教信息等服务，馆员应具备信息分析、创新、网络安全维护、数据统计挖掘能力。专业医院图书馆需要配备图书情报、医学、计算机、英语、统计分析等相关专业高素质人才。近年来交大医学院图书馆面向附属医院临床医务人员和馆员进行数据库检索培训，内容涉及 UpToDate、EBSCO 等临床循证医学数据库及 Endnote 医学统计软件、Web of Science 文献计量分析工具，开展信息化病因查找及过滤等知识讲座；同时通过参加上海市图书馆学会和中国图书馆学会医学图书馆专业委员会年会推动馆员专业能力提高。医院图书馆一方面要设法引进医学信息学专业人才，另一方面要加强馆员培训，以适应医学信息学迅速发展，为科研不同阶段提供针对性信息知识服务，推动科研质量提升。

5 结语

智慧医院图书馆能为临床医疗、医学科研教学和医院管理提供更加便捷的服务[1]。在传统医院图书馆向智慧医院图书馆转换过程中，对管理人员业务水平和综合能力提出更高要求，医院领导应引起重视，积极培养和引进图书情报高级人才，推动智慧医院图书馆形成与发展。

参考文献

1 王世伟．未来图书馆的新模式——智慧图书馆［J］．图书馆建设，2011（12）：1－5.

2 冯琦．医院图书馆精细化信息服务探讨［J］．医学信息学杂志，2016，37（11）：82－84，88.

3 中国医院协会．三级医院图书馆设施规程［EB/OL］．［2011－10－10］．http：//www.cha.org.cn.

4 陈平华．基于云计算技术的高职院校图书馆发展对策初探［J］．电子世界，2016（18）：28.

5 沈沂．医疗大数据对医院图书馆文献信息服务的影响［J］．图书情报工作，2016，60（S1）：33－36.

6 祝业，张建平，王璇，等．智慧图书馆服务平台建设的思考［J］．中华医学图书情报杂志，2018，27（6）：72－74.

7 王非．基于网络环境下的高校档案管理系统信息安全问题分析［J］．科技资讯，2015，13（2）：125.

8 Keshnee Padayachee. Taxonomy of Compliant Information Security Behavior［J］. Computers & Security，2012，31（5）：673－680.

9 陈臣．基于大数据的图书馆个性化智慧服务体系构建［J］．情报资料工作，2013（6）：75－79.

10 董瑞玉，冯占英，张晓梅，等．基于大数据应用的医学图书馆服务定位［J］．医学信息学杂志，2017，38（1）：75－78.

（上接第 56 页）

参考文献

1 Bifan Zhu，Yanfang Wang，Jian Ming，et al. Disease Burden of COPD in China：a systematic review［J］. International Journal of Chronic Obstructive Pulmonary Disease，2018（13）：1353－1364.

2 沈洪兵．大数据时代的临床医学研究——机遇和挑战［J］．南京医科大学学报（自然科学版），2020，40（3）：303－305.

3 郑劲平，简文华．慢性阻塞性肺疾病标准数据集［M］．北京：人民卫生出版社，2019.

4 黄跃，魏岚，张蕾，等．基于大数据的医院信息集成平台建设［J］．中国医学装备，2019，16（4）：103－105.

5 陈钟，关志向．国产密码体系在区块链中的应用与挑战［J］．中国信息安全，2019（11）：71－73.

6 夏旭光．光网络传输技术在电信网中的应用策略［J］．中国新通信，2019（11）：103.

7 张冬莹，高怡，简文华，等．肺功能检查技术在基层医疗卫生机构推广可行性及建议［J］．中国全科医学，2020，23（29）：3638－3643.

8 崔春舜，徐畅，高东平．国际健康大数据研究计划发展及启示［J］．医学信息学杂志，2019，40（12）：8－12.