

基于电子病历的呼吸专病医疗大数据综合应用平台数据质控体系构建及应用^{*}

赵前前

李嘉琛 梁立荣

(首都医科大学附属北京朝阳医院信息中心 北京 100020)

(1 首都医科大学附属北京朝阳医院临床流行病学研究室(烟草依赖治疗研究室) 北京 100020

2 北京市呼吸疾病研究所 北京 100020)

〔摘要〕 基于医院电子病历搭建呼吸专病医疗大数据综合应用平台, 构建全流程数据质量控制体系, 详细阐述体系构建方法, 包括数据平台基本架构、数据资源的整合与标准化处理、全流程质控体系构建、质控管理制度建立, 分析该体系应用效果, 讨论相关问题。

〔关键词〕 数据质量; 大数据平台; 真实世界研究

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.07.010

Establishment and Application of the Data Quality Control System of an Electronic Medical Record – based Big Data Platform for Respiratory Diseases ZHAO Qianqian, Information Center, Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, China; LI Jiachen, LIANG Lirong, 1Clinical Epidemiology Department (Tobacco Dependence Treatment Department), Beijing Chaoyang Hospital, Capital Medical University, Beijing 100020, 2Beijing Institute of Respiratory Medicine, Beijing 100020, China

〔Abstract〕 The whole process data quality control system is built based on an electronic medical record – based big data platform for respiratory diseases. The paper expounds the system construction method in detail, including the basic architecture of the data platform, the integration and standardized processing of data resources, the construction of the whole process quality control system, and the establishment of the quality control management system, analyzes the application effect of the system, and discusses relevant issues.

〔Keywords〕 data quality; big data platform; Real World Study (RWS)

〔修回日期〕 2021-11-04

〔作者简介〕 赵前前, 高级工程师, 发表论文 16 篇; 通信作者: 梁立荣, 博士, 研究员。

〔基金项目〕 北京市属医院科研培育项目“人工智能辅助的慢阻肺急性加重期诊疗决策支持模型的构建”(项目编号: PX2020014); 北京市科技计划项目“AECOPD 住院患者不同救治方案的效果评价”(项目编号: No. Z201100005520028)。

1 引言

基于海量电子病历数据开展真实世界研究并为临床实践提供真实世界研究证据, 已成为现代临床医学研究趋势。相比于欧美国家, 我国开展此类研究具有数据资源丰富的优势。利用专病病历资源建立医疗大数据应用平台有助于发挥医院专科特色, 将优势资源转化为学科发展的动力源。数据质量是真实世界研究数据转化为高质量研究

证据的关键，也是医疗大数据平台应用所面临的重要技术挑战。目前医疗机构存储的电子病历信息存在以多源异构和非结构化数据为主、医学术语标准不统一等问题，建立行之有效的质量控制标准和体系十分重要。国内多家医院建立专病大数据平台^[1-4]，但既往文献对数据质控过程报告比较有限，且不同研究之间存在差异，缺少公认、系统化的质控体系。2018 年北京市呼吸疾病研究所（以下简称呼吸所）和北京朝阳医院共同搭建呼吸专病医疗大数据综合应用平台。为了充分利用医院优势专科电子病历资源、促进高质量真实世界研究产出，院所联合从技术和管理制度两个层面构建全链条数据质控体系，本研究旨在分析与评价该质控体系的应用效果。

2 方法

2.1 数据平台基本架构

北京朝阳医院通过整合医院信息系统（Hospital Information System，HIS），检验信息系统（Laboratory Information System，LIS），医学影像存储与传输系统（Picture Archiving and Communication System，PACS）等临床业务系统数据，于 2014 年建立临床数据中心^[5]。在此基础上 2018 年呼吸所与朝阳医院联合搭建呼吸专病医疗大数据综合应用平台。该平台首先从临床数据中心提取呼吸系统疾病患者门诊及住院的全部医疗数据信息，经过自然语言处理、数据转换和数据清洗，形成面向数据使用者的综合应用平台，见图 1。

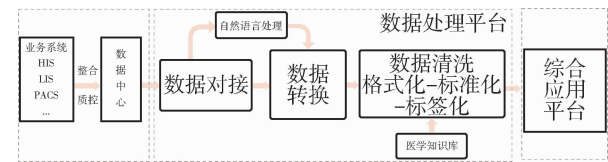


图 1 呼吸专病医疗大数据综合应用平台
整体架构和数据处理过程

2.2 数据资源的整合与标准化处理

数据的提取和处理包括以下步骤：数据中心对各临床业务系统数据进行整合和初步质控；平台数据对接，采用 Web Service 接口从数据中心获取数据，进入数据处理平台；自然语言处理，对电子病历中的非结构化文本数据采用条件随机场（Conditional Random Fields，CRF）模型进行实体抽取^[6]；数据转换，基于源数据和映射关系表将多源异构数据统一成标准的数据结构；数据清洗，包括数据格式化，基于规则表对空值、特殊符号、医疗专用名词进行格式化；数据标准化，根据临床专家建议、参照国内外标准，分别构建症状、诊断、药品、手术、检验和影像检查的医学标准知识库，将非标准的医学实体名称转换为标准的名称，见表 1；标签化，由医学专业人员将标准化的医学术语归类打标签，使数据颗粒度更好地满足临床科研应用需求，由多组人员进行交叉核验以确保标签的准确和规范化；形成面向呼吸专科医护人员结构化科研应用数据库。

表 1 医学知识库术语标准词量

类别	参考标准	标准词量
药品	国家基本药物目录	40 801
诊断	ICD - 10	45 895
手术	手术操作分类代码国家临床版 3.0	21 391
检验	临床检验项目分类与代码（wst102 - 1998）	2 322
影像检查	医学影像学（第 7 版）	2 944
症状	诊断学（第 8 版）	8 605

2.3 全流程质控体系

2.3.1 质量控制内容 参考既往文献，本研究主要考虑完整性、正确性、一致性、时效性以及合理性 5 个数据质量维度^[7]。在大数据平台建立过程中可能发生提取错误、映射错误以及源数据错误^[8]，因此在数据处理的各环节采取相应质控措施，形成质控体系，见表 2。

表 2 呼吸专病医疗大数据综合应用平台全链条质量控制内容

平台	环节	维度	主要内容
数据中心	多系统数据整合	完整性	监控从医院各业务系统获取的数据量
		合理性	统计药占比、费用和死亡人数等数据，与期望值比较
		一致性	核查是否匹配错误
数据处理平台	平台数据对接	时效性	监控数据获取时间
		完整性	监控采集的患者记录数量和各变量的缺失记录占比
	自然语言处理	正确性	抽查诊断、手术、症状、疾病、药品名称识别情况
		完整性	比较转换前后的记录数量
	数据转换	正确性	判断字段及内容映射是否正确
		完整性	比较处理前后的记录数量
综合应用平台	研究数据库质量评价	正确性	抽查格式化、标准化、标签化后的数据是否正确
		完整性	计算各变量缺失值占比
		正确性	基于研究目的进行专业复核
		合理性	判断各变量是否在特定取值范围以内，计算异常值占比

2.3.2 质量控制规则 对于质控环节中的每项内容设定具体质控条目，采用基于规则的质量评价方法^[9]，每条规则是一个逻辑判断和相应的阈值标准，检查对象为数据库、表或变量。当数据达到标准时判别为通过该项检查，如果未达到标准则记录问题并及时反馈，触发程序或通过人工进行处理。

2.3.3 质量控制流程 参考观察性健康医疗数据科学与信息学（Observational Health Data Sciences and Informatics, OHDSI）的数据质量仪表板（Data Quality Dashboard, DQD），结合前期发现的实际数据问题，构建适合呼吸专病大数据平台的质控体系，包括 5 个维度、208 条质控规则，覆盖数据整合和处理全流程。在数据中心的多系统数据整合环节，主要是保证数据获取完整、多源数据匹配正确。完整性条目是比较原始业务系统和平台获取的数据量；一致性条目是检查同一患者多次住院数据之间的基本信息（姓名、年龄和性别）一致性；此外每日统计药占比、费用和死亡人数数据与统计室比对。在数据处理平台需要保证数据采集的时效性、完整性以及数据处理的正确性。在数据接口监控数据获取时间；在每一个处理步骤设置数据量的完整性检查条目；对于非结构化电子病历的文本提取，计算精确率和召回率评价识别的正确性；对于

标准化过程，通过计算处理后的名词与标准词之间的 TF-IDF 余弦夹角相似度来评价正确性。在综合应用平台进行统计描述，计算最终科研数据库中变量的分布情况、缺失值和异常值比例。此外在实际使用数据前研究者还需基于特定研究目的和标准进行相应数据处理。使用 Python 和 SQL 编写质控程序，实现大部分条目的自动化核查，少数条目需要定期进行人工核查，如自然语言处理的识别准确率。

2.4 建立健全质控管理制度

为了确保数据质控的规范化和常态化建立规章制度。每日质控程序自动核查所采集数据，由平台专职人员在各环节对数据质量进行实时监控，第一时间发现并记录问题，根据问题原因向相关人员反馈、及时解决，必要时更新质控程序，纳入新的质控条目；每周定期和数据中心比对数据量，确保一致；每日导出数据质控报告，汇总数据情况以及存在问题；建立数据字典，保留质控过程文档，明确记录数据处理方法。医院病案科会核实病案首页的诊断编码并随时更新电子病历，数据中心每周会重新抽取数据，更新数据库。质控团队人员均严格执行上述各项制度，以保证每个环节的数据质量满足应用需求。

3 结果

3.1 平台收集数据情况

北京朝阳医院的呼吸专病医疗大数据综合应用平台目前已积累 2013 年 8 月 - 2021 年 4 月的临床数据，其中住院 146 745 人次，门诊 109 615 人次，共有 5 222 个病种。数据内容包括患者基本信息、疾病史、检查、检验、用药、手术以及诊断。

3.2 数据中心一致性核查结果

数据质控系统在数据处理各环节进行不同维度和层次的检查，检查结果达到预先设定的阈值才被认为合格。数据中心一致性核查结果显示，同一患者多次住院数据的姓名、出生日期、性别、民族等基本信息的一致率高，在现有住院数据中只发现 1 例姓名不匹配、2 例出生日期不匹配。时效性检查通过实时监控接口的数据获取时间，确保数据中心 95% 以上的数据可以及时对接到平台，见表 3。

表 3 部分质控条目检查结果示例

环节	维度	层次	规则	阈值	检查结果	是否合格
多系统数据整合	完整性	表	基本信息表记录数 = 业务系统记录数	是	是	是
多系统数据整合	完整性	变量	病案首页出院日期非缺失记录数/总记录数 > 阈值	0.95	0.99	是
多系统数据整合	一致性	变量	同一患者两次住院信息中出生日期相同的比例 > 阈值	0.95	1	是
数据对接	时效性	数据库	未超时次数/总次数 > 阈值	0.95	0.97	是
数据对接	完整性	变量	病案首页入院时间变量非缺失记录数/总记录数 > 阈值	0.95	1	是
数据对接	合理性	变量	检验时间 < 出院时间的记录数/总记录数 > 阈值	0.95	1	是
自然语言处理	正确性	变量	抽样核查主诉症状的精确率 > 阈值	0.90	0.96	是
自然语言处理	正确性	变量	抽样核查主诉症状的召回率 > 阈值	0.90	0.96	是
数据转换	完整性	表	病案首页转换后记录数 = 转换前记录数	是	是	是
数据转换	完整性	变量	病案首页证件号码变量转换后非缺失记录数/总记录数 > 阈值	0.95	1	是
数据清洗	完整性	表	住院检查格式化后非缺失记录数 = 格式化前记录数	是	是	是
数据清洗	正确性	变量	抽样核查住院检查项目标准化的准确率 > 阈值	0.90	0.93	是
研究数据库质量评价	合理性	变量	BMI 在 10 ~ 80kg/m ² 范围内的记录占比 > 阈值	0.95	1	是
研究数据库质量评价	完整性	变量	吸烟状态变量非缺失记录占比 > 阈值	0.95	0.99	是

3.3 变量正确性核查结果

在数据处理平台核查变量正确性。为了评价自然语言处理对电子病历文本的识别效果，共抽取 3 万余份电子病历进行验证，结果总体精确率和召回率分别达到 95.45% 和 95.3%，吸烟饮酒史、症状、药品等信息的精确率和召回率均超过 96%，识别比较准确。对提取的实体名称进行标准化处理，通过计算处理后词语与标准术语之间的 TF - IDF 余弦夹角相似度来评价标准化过程的正确性。结果显示有 83.2% 的词语经过处理完全转化为标准术语（相似度 = 1），9.4% 的词语与标准术语相似度在 0.8 ~ 1 之间，没有相似度低于 0.6 的词语。在各类项目中手术、诊断和药品名称的标准化程度相对较高，而

影像检查结果的标准化程度略低，见表 4。

表 4 数据处理过程中的正确性评价

内容	自然语言处理正确性 *		标准化正确性
	精确率（%）	召回率（%）	相似度 ** > 0.8 的名词占比（%）
基本信息	-	-	-
吸烟饮酒史	96.2	96.2	-
症状	96.2	96.1	-
疾病诊断	95.2	95.2	97.0
药品	96.5	96.9	97.0
检验	-	-	93.1
影像检查	96.4	95.5	81.2
手术	95.5	95.3	100.0

注：* 表示对病历进行实体抽取，精确率为正确的实体数量/得到的所有实体数量，召回率为正确的实体数量/标准答案的所有实体数量；** 表示处理后词语与标准术语之间的 TF - IDF 余弦夹角相似度，越接近 1 代表相似度越高。

3.4 主要变量缺失率计算结果

对经过质控的科研数据库进行描述性分析, 计算主要变量缺失率: 年龄、性别、住院费用、住院天数以及出院诊断等信息均无缺失; 吸烟和饮酒缺失率较低, 分别为 0.55% 和 3.01%; 在体格检查指标中, 血压 (0.08%) 和脉搏 (0.02%) 等生命体征变量缺失率极低, 非临床必要变量身高 (44.4%) 和体重 (49.6%) 缺失率较高; 疾病史信息没有缺失。

4 讨论

4.1 概述

近年来电子病历等医学信息系统被广泛使用, 医疗大数据已成为重要临床研究资源。但常规诊疗数据不是针对科研问题专门收集的, 需要经过严格数据提取、整合以及清理过程才能形成符合科研要求的数据库。美国食品药品监督管理局针对临床研究数据提出了可溯源性、易读性、同时性、原始性、准确性、完整性、一致性、持久性、可用性的要求, 即 ALCOA + CCEA 原则^[10]。我国有研究者针对利用健康医疗数据开展真实世界研究制定了通用技术规范, 提出应从既有数据本身和研究型数据库构建过程两方面进行质量评价^[11]。

4.2 医疗大数据质量控制主要问题

随着真实世界研究证据的快速增加, 医疗大数据质量愈加受到重视。目前医疗机构通常建有数据中心, 即整合各临床业务系统形成的数据库, 但由于缺乏实际应用, 往往未实施严格的质控措施。本研究在前期科研应用中发现许多数据质量问题, 包括源自数据中心的数据缺失、关联错误, 以及源自原始数据的术语不标准、信息不完整等。因此医院和呼吸所联合构建质控体系, 从数据中心质控开始, 重在控制数据缺失和多源数据匹配错误; 对呼吸专病大数据综合应用平台的数据对接、数据转换和数据清洗各环节进行长期持续质控; 同时建立反馈机制, 根据临床医护人员在应用中发现的新问

题, 不断查找质控漏洞、完善质控环节、细化质控条目和标准, 从而构建临床应用驱动的全链条、动态完善的质控体系, 并以规章制度确保各项质控措施落实常态化与标准化。

4.3 数据质量控制标准问题

为了使质控体系具有科学性, 本研究参考了既往数据质量的理论框架。数据质量涵义复杂, 涉及多个不同维度, 而且存在不同术语和概念^[12]。有研究者归纳出了完整性、正确性、一致性、时效性以及合理性这 5 个电子病历数据质量维度^[7]。不同维度之间可能存在一定重叠, 如合理性常作为正确性的代理指标, 通过检查不合理数值识别数据错误。基于质量维度的理论框架产生了具体的评价指标和工具。OHDSI 设计了通用数据模型 (Common Data Model, CDM)^[13], 以统一标准对临床数据进行抽取、转换、加载 (Extract - Transform - Load, ETL), 并开发配套质量评价工具 ACHILLES 和 DQD。但这些质控工具对原始数据的标准化程度要求较高, 难以直接应用到国内^[14]。本研究从实践出发, 将理论与现实数据问题相结合, 探索性地在国内一家综合医院构建了系统化医疗大数据平台质控体系。但国内不同医院之间的临床信息系统存在差异, 原始数据质量参差不齐, 未来还需加强国内医院的信息化建设, 统一标准, 以实现医疗大数据应用的互联互通。

4.4 建立通用数据标准

专病数据库可以有力支持医院特长专科临床研究, 我国已有多家医院建立了专病大数据平台, 并依托平台数据开展研究^[1,3,15]。在提取自电子病历的数据库中, 临床专业术语往往存在不规范、不统一的问题, 因此建立通用数据标准十分重要。本研究发挥北京呼吸疾病研究所的学科优势, 参考国内外现有标准建立呼吸专科医学知识库, 对电子病历中的专业名词进行了标准化, 使研究者可以通过标准术语进行数据查询和筛选, 提高科研数据质量。今后可以将呼吸专科数据标准进一步完善、推广, 建立区域、国家层级的通用标准, 推动不同医院间

的数据共享和科研合作。

5 结语

本研究基于呼吸专科医疗大数据综合应用平台的科研需求,构建了从多源数据整合、平台对接、自然语言处理、数据转换、数据清洗直到最终综合应用平台的全流程质量控制体系。应用后数据库中重要临床相关变量缺失率降低,检验、检查、药品、诊断等医学术语基本实现标准化,可以支持呼吸专科医护人员开展呼吸系统疾病诊治相关临床研究。

与既往质量评价系统相比,本研究构建的大数据平台质控体系的主要特点在于全流程动态监控。针对数据整合、对接、转换和清理各环节存在的主要问题设置相应质控条目,在每日获取数据的同时动态监测数据质量。考虑到数据质量问题往往产生于实践中,本研究以制度确保质控工作常态化,在科研应用中逐渐扩充质控条目体系。但现有质控措施主要解决数据提取和转换过程中发生的各类错误,而不是临床数据质量问题,例如病历书写不规范导致关键信息缺失或不准确等。质控系统有助于识别此类问题,通过反馈促使临床医生提高病历书写质量,增强标准化程度。今后可以通过建立专科电子病历系统,特别是完善研究型病房住院电子病历系统,从根源上加强原始数据填写的规范性,确保后期应用于真实世界研究的数据质量。目前本研究构建的全流程质控体系主要适用于呼吸专科疾病的住院数据资源,未来可以和更多学科专业组对接,将医学标准知识库进一步扩展,建立并实施适合综合医院的多学科住院数据资源质控体系,这对促进综合医院临床数据资源的转化应用及学术影响力的提升具有现实意义。

参考文献

- 1 刘景丰,刘红枝,陈振伟,等.肝病和肝癌大数据平台建设体系及其初步应用[J].中华消化外科杂志,2021,20(1):46-51.
- 2 方莹,陈智,简文华,等.呼吸专科数据平台多中心数据集成与应用[J].医学信息学杂志,2021,42(2):53-56,83.

- 3 林琳,王韬,甘伟,等.基于电子病历系统的脑血管病专科大数据科研平台设计与应用[J].中国卒中杂志,2020,15(6):616-619.
- 4 郭强,王丛,衡反修.医疗大数据平台建设机遇、挑战及其发展[J].医学信息学杂志,2021,42(1):1-8.
- 5 卞松,刘宇宏,高黎,等.以数据挖掘为核心的医院数据管理平台建设初探[J].中国病案,2015,16(1):39-42.
- 6 梁立荣,李长伟,沈晔,等.基于层叠条件随机场模型的电子病历文本信息抽取[J].计算机应用与软件,2019,36(10):47-54,112.
- 7 Weiskopf N G, Weng C. Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research [J]. J Am Med Inform Assoc, 2013, 20(1): 144-151.
- 8 Huser V, Defalco F J, Schuemie M, et al. Multisite Evaluation of a Data Quality Tool for Patient - Level Clinical Data Sets [J]. Generating Evidence and Methods to Improve Patient Outcomes, 2016, 4(1): 1239.
- 9 Wang Z, Dagtas S, Talburt J, et al. Rule - Based Data Quality Assessment and Monitoring System in Healthcare Facilities [J]. Studies in Health Technology and Informatics, 2019(257): 460-467.
- 10 李雪迎,沙若琪,姚晨,等.面向真实世界数据的临床研究数据治理模式选择[J].中国循证医学杂志,2020,20(10):1150-1156.
- 11 王雯,高培,吴晶,等.构建基于既有健康医疗数据的研究型数据库技术规范[J].中国循证医学杂志,2019,19(7):763-770.
- 12 Kahn M G, Callahan T J, Barnard J, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data [J]. Generating Evidence and Methods to Improve Patient Outcomes, 2016, 4(1): 1244.
- 13 Stang P E, Ryan P B, Racoosin J A, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership [J]. Annals of Internal Medicine, 2010, 153(9): 600-606.
- 14 孟若谷,杨羽,张路霞.健康医疗大数据质量评估方法进展与展望[J].中国卫生信息管理杂志,2019,16(6):677-681.
- 15 戴静娟,胡萍,陈名扬,等.耳鼻喉科专科大数据平台的设计与实践[J].中国卫生信息管理杂志,2021,18(3):405-410,425.