doi:10.3969/j.issn. 1672-5166.2019.05.04

上海市宝山区重点慢病管理数据质量控制研究

薛俊磊[①] 胡利娟[②] 吴 萃[①△]

文章编号：1672-5166（2019)05-0537-05 中图分类号：R-34；R319 文献标志码：A

摘 要 为解决区域卫生数据质量可信性问题，提高信息化应用水平，根据宝山区区域信息平台慢病管理数据质量控制的工作实践，通过对数据完整性、准确性、可信性和一致性的全面评估及质量控制举措成效分析，提出了区域数据质量控制体系建设思路。

关键词 区域卫生信息化 数据质量评估 控制策略

Research on Quality control of Key Chronic Disease Management Data in Baoshan District of Shanghai

XUE Junlei, HU Lijuan, WU Cui

Shanghai Municipal Center for Diseases Control and Prevention, Shanghai 201901, China

Abstract Data quality is the key factor to determine the application level of informationization. The credibility of regional health data quality is an urgent problem to be studied and solved. Based on the practice of data quality control of chronic disease management in Baoshan District Regional Information Platform, combining with a comprehensive assessment of data's saturation, accuracy, credibility, consistency, and the analysis on effect of quality control measures, this paper puts forward the constructive idea of regional data quality control system.

Keywords regional health informatization; data quality evaluation; control strategy

0 引言

目前，我国正处于信息化高速发展的时期，医疗卫生数据正在步入大数据时代。2016 年 6 月，国务院办公厅印发了《关于促进和规范健康医疗大数据应用发展的指导意见》。2016 年 10 月，中共中央、国务院印发了《“健康中国 2030”规划纲要》。这些文件明确提出了我国应加快推进健康医疗大数据的挖掘和有效利用，使卫生信息化在提高居民医疗健康服务质量和自我健康管理能力上发挥重大作用，并为前瞻性卫生决策提供依据。上海市宝山区基于健康档案的区域卫生信息平台（以下简称“EHR 平台”）建设以居民电子健康档案为基础，通过整合居民健康管理及医疗信息资源，实现公共卫生、计划生育、医疗服务、医疗保障、药品

基金项目：上海市疾病预防控制创新特色项目；宝山卫生青年医学人才培养计划（项目编号：bswsyq-2016-A22）

① 上海市宝山区疾病预防控制中心，上海市，201901

② 上海市淞南镇社区卫生服务中心，上海市，200441

作者简介：薛俊磊（1986-），男，本科，主管医师；研究方向：疾病控制；E-mail:raychar@163.com

通信作者：吴萃（1981-），女，本科，主管医师；研究方向：疾病控制；E-mail:25850569@qq.com

△ 通信作者

供应、综合管理 6 大业务应用系统的资源共享和业务协同，促成健康相关医疗信息快速汇集，档案份数及数据体量呈指数级快速增长。以电子健康档案为主要载体的医疗健康数据广泛应用于临床决策支持、药物研发、远程病人数据分析、公共卫生领域等方面，发挥巨大价值[1-3]。

经过多年的系统应用实践，宝山区 EHR 平台已汇集了海量包括高血压、糖尿病在内的重点慢病管理数据（以下简称“管理数据”），涵盖了贯穿于全生命周期的健康教育和健康促进、健康行为和生活方式养成，强化慢性病早期筛查和早期发现等主要职能。完整、及时、准确的数据是实现信息平台职能、推动慢性病防控模式转变，利用信息采集技术辅助疾病干预和提高公众健康自主管理能力的基础。然而，在社区管理效果评估及实际考察中发现仍然存在一定的问题，数据无法反映本地健康管理的真实效果。EHR 平台面临的错误数据、脏数据和数据可信性度低等一系列数据质量问题[4-5]，已普遍成为限制区域卫生数据深度挖掘、系统效能发挥和卫生决策分析的瓶颈。本研究旨在通过对辖区 EHR 平台管理数据的质量评估和治理实践，探索发现数据问题产生原因，并提出保障数据质量的长效机制。

1 区域数据质量控制

高质量的数据应该是能充分满足用户利用要求的数据。由于使用场景和应用需求不同，数据质量的概念主观性较强，不同的系统用户对数据质量的评价标准并不相同。对卫生信息数据质量评估体系的维度指标选择应满足用户不同的评估决策需求，并兼顾数据获取的可行性和可操作性。本研究基于辖区业务监管需要和考核评估标准实施需求分析，并结合 EHR 平台应用场景设计创建数据质量控制模型，评估规则涵盖了评价电子健康档案数据的主要维度，包括数据的精确性、可信度、关联度和安全性、可用性和实效性等[6]。通过数据质量控制实践综合评价区域卫生信息数据质量和数据挖掘有效利用的可行性。

1.1 完整性评价

完整性即数据完整程度，是数据评估可信和结果可利用的基础，不仅是信息生产需求也应满足目标用户需求。即不仅要求管理数据避免出现核心字段和重要字段的缺失，同时应在管理频次和规律上符合业务管理规则。本研究全量提取了辖区管理数据个案全业务流程信息实施完整性质控，通过非空字段数据量与全数据量的比值计算字段总体饱和度[7]，并基于业务规则核查管理流程规范性，综合评估系统数据完整度和可利用度，得到宝山区慢病管理数据完整性评价。评估内容包括健康档案个人信息核心字段缺失情况，首访记录重要字段采集情况，随访数据字段缺失情况以及管理评估动作的完整性和规范性等。

1.2 准确性评价

准确性用以评价管理数据的正确性和精确度。要求管理数据的字段格式、长度、取值、编码等不仅应符合区域信息平台校验标准，也应符合业务管理要求的统一标准。通过确定管理数据的值域标准和逻辑关系规则，全量提取字段评估数据值域可疑问题，综合评价管理个案准确性。纳入评价的值域标准基于自然规律、社会常识和业务管理规范等确定，如身高、体重、收缩压、舒张压、实验室检测结果等应在合理取值范围。逻辑关系规则反映了数据之间是否存在约束的保障，以及相互关联关系的合理性，包括个案维度和管理维度：要求个案信息内部数据间符合关联逻辑合理性，并且病例确诊、纳入管理、接受随访等具有时序特征数据的信息应满足时效性。

1.3 可信性评价

由于 EHR 平台的慢病管理数据采取自疾病管理的各个环节，数据可信性的内涵既包括个体特征标签，也包含疾病管理的时序特征。通过分析管理数据在个体和群体不同维度的分布规律，通过个人健康评估数据的可疑分布和管理情况分布合理性两部分多个评估指标，评价慢病管理数据的合理性和可信性。其中，对个体维度

而言，涵盖了全周期的健康数据评估结果应遵从时序合理性。例如在管理时序上，个案信息不应出现血糖血压等实验室指标长期无变化或身高出现异常波动的情况。对群体维度而言，数据应符合群体特征和统计学规律。例如疾病临床症状、并发症发生、核心指标控制情况等计数资料分布应遵从区域流行状况；体格检查和实验室指标等计量资料分布的集中趋势和离散趋势应符合一般统计学规律；管理频次、随访时间安排和方式选择等管理信息不应呈现明显的偏态分布等。

1.4 与强可信数据源的一致性评价

1.4.1 市区两级平台数据比对

EHR 平台运行过程中，用户会因为实际使用需求与设计功能规范不一致，而执行一些删除、修改等造成数据破坏的操作。同时在长期运行维护过程中，也会由于数据恢复、更新不及时或多个事务冲突访问等情况，造成问题数据或“脏数据”的产生。通过实施区级全量管理数据与经校验成功上传上海市级平台数据的比对，排查上传数据和本地应用数据的差异及可能出现的其他质量问题。评估数据传输更新情况、抗篡改能力和脏数据比例。

1.4.2 区域平台大数据关联一致性

应用大数据技术，在区域卫生平台数据仓库内抽取源于 HIS 系统的疾病诊断信息、用药信息，源于 LIS 系统的实验室检查结果和源于疾病直报网络的报病信息和死因统计信息等强可信数据作为依据。通过与区域管理数据的关联和比对发现不一致问题，确定患病情况、用药情况、诊疗服务利用、随访管理开展及核心指标控制情况可疑数据，实施多维度数据质量评价。

2 应用情况

2.1 质控结果

2018 年度宝山区全量提取 EHR 平台本地管理数据 6 721 580 条与经校验成功上传市级平台管理数据 9 057 444 条，组织实施两级全量数据对比和区域数据质量控制评估。通过质控共排查发现各类错误数据 2 496 778 条，其中包括市级平台重复数据 2 436 947 条（占市级数据总量 26.90%）和区级平台其他错误数据 59 831 条（占本地数据总量 0.89%）。并在区域数据质量评估中发现辖区内管理数据存在不同程度上的准确性、可信性和一致性问题，数据可疑是本区各管理机构存在的共性问题。主要表现为数据源真实性可靠性较差，数据 ETL 过程中产生大量历史遗留错误，市、区两级平台的数据体量不匹配，数据维护和监管缺位造成问题扩散和放大等。

2.2 问题分析

宝山区区域卫生数据中心整合多系统多库建立 EHR 业务数据库，所纳入的医疗卫生数据具备多源性、异构性、离散性和值域分布复杂性的特点。因此，管理数据在数据源和抽取、转换、加载（ETL）各个环节中均可能产生各种问题[7]。

2.2.1 数据源

在数据生产环节中业务管理人员由于对信息规则规范的不熟悉，存在数据录入不准确、不完整或不规范等问题。或在遇到系统问题时，部分缺乏责任心的人员执行了违规的修改或删除数据操作。同时，系统软件缺陷或运行错误也会引入问题数据。这些因素直接造成了数据源产生部分错误数据。

研究发现宝山区 EHR 平台在应用更新和数据运行维护过程中出现了市区两级数据体量的明显差异，其中由于主键约束的错误变更造成的大体量数据错误重复上传，严重影响了数据的业务应用和管理效用的发挥。事实上，在质量评估的过程中发现，由于约束性的下降，造成了不良数据的错误添加，进而在运行传输环节引起连锁反应，最终产生了生产平台和应用平台间的大体量数据差异。因此，在平台设计和运行环节，保障主键作为“标识”属性的约束性，对于维护数据的有效性和一致性具有重大意义，尤其应该在平台建设、更新和维护环节中得到应有的重视。

2.2.2 ETL 过程

区域 EHR 平台是个逐步建设整合的过程，不可避免地存在频繁的 ETL 过程。宝山区 ERH 平台数据汇集

和应用的过程中，经历了各基层医疗机构分散管理、区级数据规范整合和市区两级应用平台统一部署等几个关键时间节点。在建设初期缺乏统筹规划，建设进程中业务系统功能规范不断修正，整合阶段新系统规范要求与历史系统设计冲突等因素影响下，ETL 过程存在较多错误隐患。全国各地 EHR 平台建设过程中也存在类似的实施经验：承建信息厂商为解决由于接口库约束过多导致的无法上传或整合数据问题，接口库去除主外键约束、非空约束、字典表约束方便系统建设等[8]。缺乏约束的 ETL 过程中出现字段标准、字典码映射和抽取规则的不统一、不准确时，这些问题未能被及时发现，均造成隐匿的原始数据错误产生和留存。

2.3 改进措施

依托于区卫健委，以数据仓和模型库等相关技术为支持，宝山区已设计建立适用于区域需求的数据质控模型，通过技术手段和管理机制并进，初步建立区域慢病管理数据质量评估体系。质控模型设计包括数据完整性、准确性、可信性和一致性 4 大维度，27 类分级指标，共 113 项质控内容，见表 1。并以数据评估质量为依据，在区域卫生机构考核管理中落实奖惩机制。2018 年度开始，通过在卫生信息平台中应用数据质控模型，引入强可信数据关联校验，形成区域质控机制，定期开展数据质量评估并落实整改，宝山区重点慢病管理数据质量得到极大的改善，以核心指标为例，数据整体可信度得到了较大的提高，见图 1。

通过建立对数据的全生命周期的追踪监控机制[7]和管理手段，强化数据运维职责，监控数据采集全流程，包括数据从生产、抽取、转换、上传等全过程的跟踪信息，及时掌握数据 ETL 过程中发生的问题，较快地进行问题划界与定位，及时交付业务人员核实修正。

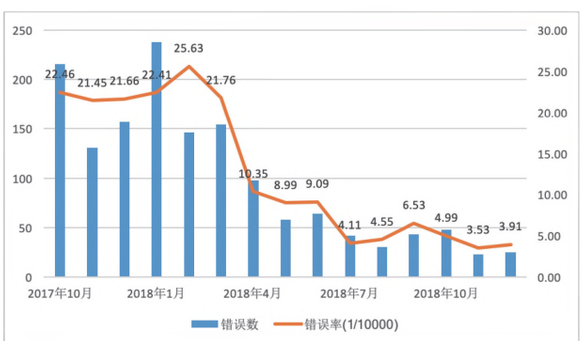

图 1 2017 年 10 月至 2018 年 12 月管理数据与死亡库比对情况

表 1 宝山区慢病管理数据质量评估体系分级指标分布

维度	分级指标	质控项目
数据完整性	1.1 个人信息完整性	4
	1.2 首访评估记录完整性	4
	1.3 随访记录核心字段完整性	6
	1.4 管理评估动作完整性和规范性	3
数据准确性	2.1 疾病自然史合理性	3
	2.2 重复数据核查	2
	2.3 管理数据取值合理性	9
数据可信性	3.1.1 随访方式合理性	4
	3.1.2 随访管理安排规范性	2
	3.1.3 随访间隔分布	1
	3.1.4 个人实验室指标时序分布波动	2
	3.1.5 疾病相关症状分布	1
	3.1.6 重点指标控制情况	4
	3.2.1 规范管理情况	3
	3.2.2 非活跃数据分布	2
	3.2.3 随访方式及随访量分布	3
	3.2.4 随访者单日随访次数分布	2
	3.2.5 重点体格检查指标取值偏好	2
	3.2.6 重点实验室指标群体分布趋势	9
	3.2.7 干预内容群体分布	1
强可信数据源一致性	4.1 共病患者管理数据相互校对	7
	4.2 与死亡数据校对失访情况	2
	4.3 与诊疗数据校对管理情况	2
	4.4 与检验数据校对重点指标控制情况	2
	4.5 与用药数据校对服药明细	2
	4.6 与诊疗数据校对心脑事件发生情况	7
	4.7 与诊疗数据校对并发症发生情况	24

3 讨论

3.1 正视数据质量现状

宝山区慢病管理数据整体可信度和可利用率低于预期。如何解决数据利用需求和数据质量管理要求，必须立足业务需求，设计符合应用实际的质量控制模型，跟进质量监管机制，通过技术手段和管理手段协同，建立了全面系统的区域数据质量控制体系。通过本研究的数据质量控制实践，区域数据质量上取得较大的改善。

3.2 完善数据评估体系

有别于传统模式，本研究广泛应用大数据技术手段，充分挖掘区域平台的强可信数据源，通过关联比对，综合评估管理数据准确性、可信性等多维度的质量。相比于费事费力且抽查比例较低的人工核查模式，强可信数据源的利用和关联，大大提高了质控覆盖面和实施效率。完善的质控体系不仅有助于数据问题的早发现早评估，保障区域平台数据质量，也促进辖区健康数据互联互通、信息共享。高效率的数据评估和高质量的数据利用在卫生行政部门的资源管理、政策制定等方面具有一定的指导意义[9]。

3.3 提高生产系统管理效能

必须注意到评估具有滞后性，单纯依靠数据质控模型的定期评估并不能实时发现所有问题。事实上业务管理机构数据维护缺位或交付供应商处理的情况普遍存在，虽然通过管控手段推进已有所改观，但大部分数据问题处理耗时仍然较长。因此，信息平台建设应兼顾应用机构的自查手段，方便非专业人员进行数据自查，例如提供统计指标与统计数据的实时反馈，通过明细数据与统计指标的比对，在数据质量提升上可获得较好的成效[8,10]。

3.4 构建全周期监管机制

数据治理需从数据源头抓起，加强从产生到应用整个生命周期的全过程监管与考核[7]，强化业务应用和ETL过程中的环节检验，才能使数据质量得到根本改善。追求短期速效的治理策略无法达到根本治理的效果。数据治理过程会催生新的流程或新的工作模式，发现新的规律，产生新的规则，继而进行流程的再评价和再监督，是一个不断运转不断形成正向的反馈的闭环过程[11]。管理部门应把质量控制评价融合到医疗业务全过程中，坚持开展数据治理工作，并形成惯性运行。■

参考文献

[1] 张国明，陈安琪．基于区域健康信息平台的医疗大数据利用探索[J]. 中国卫生信息管理杂志，2016, 13(3): 290-294.

[2] 杜明超，洪建，颜雨春，等．健康医疗大数据的应用范围与价值分析[J]. 中国卫生信息管理杂志，2017, 14(5): 652-654.

[3] Cowie M R,Blomster J I,Curtis L H,etal.Electronic health records to facilitate clinical research[J].Clinical Research in Cardiology, 2017, 106(1): 1-9.

[4] 龙虎，邱航，吴沧浪，等．四川省健康医疗大数据中心构建探讨[J]. 中国卫生信息管理杂志，2017, 14(1): 15-18,23.

[5] 孟群，毕丹，张一鸣，等．健康医疗大数据的发展现状与应用模式研究[J]. 中国卫生信息管理杂志，2016,13(6): 547-552.

[6] Weiskopf N G, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research[J]. Journal of the American Medical Informatics Association, 2013, 20(1): 144-151.

[7] 张国明．区域人口健康信息平台数据质量控制[J]. 医学信息学杂志．2017, 39(4): 67-70.

[8] 孔斌，蔡佳慧，宗文红．数据质量控制在区域卫生信息平台的实践与思考[J]. 中国卫生信息管理杂志，2014, 11(2): 169-173.

[9] 董雪，夏晨曦．区域卫生信息平台的有效性评价指标体系构建[J]. 中华医学图书情报杂志，2017, 26(8): 17-24.

[10] 崔欣，曹剑峰，陈雯，等．医疗大数据与统计数据的差异分析及应用思考[J]. 中国卫生信息管理杂志，2016, 13(6): 632-634.

[11] 费晓璐，李嘉，黄跃，魏岚，等．医疗大数据应用中的数据治理实践[J]. 中国卫生信息管理杂志，2018, 15(10): 554-558.

[收稿日期：2019-08-10　修回日期：2019-08-31]