

区域人口健康信息平台数据质量控制

张国明

(1 江苏省卫生统计信息中心 南京 210008

2 南京大学计算机科学与技术系 南京 210023)

〔摘要〕 分析人口健康信息平台数据质量存在的问题及问题产生的原因,从评价模型、数据质量评价、过程追踪和反馈几方面构建数据质量控制和评价体系,为提高平台数据质量提供保障,从而夯实健康医疗大数据共享应用基础。

〔关键词〕 区域人口健康信息平台;数据质量;评价体系

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.04.015

Data Quality Control of Regional Population Health Information Platform ZHANG Guo-ming, 1Jiangsu Center of Health Statistics and Information, Nanjing 210008, 2Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China

〔Abstract〕 The paper analyzes the problems of data quality of the population health information platform and the causes, constructs the data quality control and evaluation system from the aspects of evaluation model, data quality evaluation, process tracking and feedback, provides guarantee for improving data quality of the platform, and thus lays a solid foundation for sharing and application of big data of health care.

〔Keywords〕 Regional population health information platform; Data quality; Evaluation system

1 引言

目前,我国正处于医疗信息化快速发展的时期,国家管理部门提出建立国家、省、地市和县 4 级人口健康信息平台,有效整合和共享全员人口信息、电子健康档案和电子病历 3 大数据库资源^[1-2]。在此背景下,基于居民健康管理的区域人口健康信息平台建设和应用迅猛发展,为实现区域卫生信息共享奠定了良好基础。目前江苏省级人口健康信息平台已建立了医疗信息、健康档案、注册资源 3 大数据库,接入了 12 家省属医院、13 家市级平台、

28 家区县平台,平台数据保有量达 2 TB,数据记录数 2.2 亿多条。

随着区域平台建设的推进和业务的开展,数据质量问题已成为区域医疗业务协同、数据共享交换、健康医疗大数据分析的主要瓶颈。高质量的数据是推进健康医疗大数据融合共享的基础和分析应用的关键^[3],因此区域平台在与医疗机构或其他平台的数据对接中,建立完善的数据质量控制体系,保障数据完整、正确、及时的共享交换是迫切需要解决的问题。鉴于此,本文在分析数据质量存在问题及产生原因的基础上,从 7 个维度提出了一套数据质量评价模型和评分体系,通过建立数据质量过程跟踪和反馈机制,使区域健康信息平台的数据质量问题得到了改善。

〔修回日期〕 2016-11-11

〔作者简介〕 张国明,博士研究生,工程师,发表论文 5 篇。

2 数据质量存在的问题

2.1 不完整

医院信息系统起步较早, 早期相关标准不够完善, 平台进行数据采集时会发生记录缺失、字段信息缺失、记录不完整等情况, 最明显的例子就是患者身份证号缺失, 导致个人的医疗信息难以归集, 无法形成完整的健康档案。

2.2 不一致

主要是由于医院业务系统之间或者功能模块之间记录、编码、引用不一致等造成的。例如, 平台中同一张表的数据来自医院信息系统、电子病历等多个系统, 可能会带来平台数据主从表无法关联的问题; 又如手术编码在手麻系统、医院信息系统及住院病案系统中都存在, 但医院内部系统之间的数据不完全一致, 导致数据无法准确抽取。

2.3 错误

主要是指数据中存在错误, 如数据类型错误、范围越界、违反业务规则、逻辑错误等。

2.4 不标准

数据标准是数据共享和应用的基础。江苏省虽然制定了区域健康信息平台互联互通功能规范以及几十项数据字典, 建立数据字典的动态维护机制, 但在数据对接中仍存在大量不符合标准的数据, 如药品编码值域、空值、诊断编码错误等。

2.5 不及时

由于网络问题、采集程序问题、人为因素等, 数据有时不能及时上传到平台, 在补传数据时也可能出现漏传、错传现象。

3 数据质量问题产生的原因

3.1 概述

区域人口健康信息平台数据主要包括医疗业务

和公共卫生两部分, 包含 60 多个大类、100 多个小类数据, 覆盖 4 000 多项数据属性, 这些数据来自于医院众多业务系统, 涉及信息系统中表的数量多达 200 个, 数据值域五花八门, 标准值域少则 1 万条、多达几万条。医疗数据的多源性、异构性、分散性、复杂性等特点导致健康医疗数据在数据源和抽取、转换、加载 (Extract - Transform - Load, ETL) 过程中都容易引入各种问题。

3.2 数据源问题

3.2.1 标准规范不统一 医院在信息化建设初期缺乏计划, 未统筹规划, 导致系统仅基本可用, 缺少统一的标准规范, 同样的指标在不同系统中的定义并不相同。同时各类目录的项目和字典建立不规范, 对照不准确、不详尽、不及时调整和修改, 一些操作人员不按规范操作, 随意输入和改动项目名称和编码, 为数据标准统一带来困难^[4-5]。标准规范的不统一, 使得多个业务系统无法整合, 从而导致患者信息难以关联和汇聚, 数据的一致性就难以保证。解决此类问题的一种思路是医院建设数据中心及医院平台, 对数据字典等基础数据进行统一管理, 以患者为中心有效整合各业务系统, 从而保证数据的规范性和一致性。

3.2.2 数据录入不正确 医院业务系统众多, 由于业务系统使用人员的素质和水平参差不齐、操作不当等原因, 在录入数据时会产生各种错误, 如数据录入不准确、不完整、不一致、不规范等^[6]。解决此类问题需要强化业务系统使用人员的操作能力和工作责任心, 建立健全数据审核监督机制。

3.2.3 业务系统软件错误 在业务系统长期运行过程中, 由于软件存在缺陷很可能会引入一些错误数据。这种情况并不少见, 如果发现问题需要对系统进行修正。

3.3 ETL 过程^[7]

3.3.1 抽取错误 从源系统中抽取数据时, 一般要编制数据抽取代码或者借助一些工具配置进行, 在此过程中可能会出现编码错误或者工具配置不当, 导致原始数据正确而抽取出来的数据不正确。

此外，同一个数据项（如性别）可能存在于医院的多个业务系统中，如果医院基础数据没有统一标准，在抽取数据时必须谨慎地选择数据来源，否则可能导致数据不一致或其他错误。

3.3.2 转换错误 数据转换时需要将抽取的原始数据根据字典码表进行映射转换成符合平台标准的格式，在转换过程中可能由于数据源或转换程序问题导致错误。对部分转换错误可以通过值域校验、正则校验等校验方式检查数据的合理性，但有的是软件校验难以发现的，因此需要将软件自动校验与人工核查相结合。

3.3.3 加载错误 数据转换完成后要进行数据加载，在系统运行时可能会出现重复加载或者加载失败的情况，导致数据量异常。

4 数据质量评价体系

4.1 评价模型

4.1.1 数据范围 评价接入机构上传数据的业务范围，得分为实际上传业务与应上传业务的比值。考虑到机构的类型、规模、等级等因素，不同机构可以上传的业务范围是不同的，如三甲医院应该上传的业务要比二甲医院、社区卫生服务中心等范围广，市县平台上传数据的业务范围与医疗机构也存在差异，因此需要根据实际情况为每个接入机构配置该机构应该上传的业务范围，计算得分时也需要根据该业务范围进行。

4.1.2 业务上传率 接入机构每天实际上传数据量和应传数据量的比值，评价机构是否有数据错传、少传的情况。

4.1.3 字段饱和度 评价上传数据的完整程度，根据非空字段数据量与全部数据量的比值计算得出字段饱和度。考虑到必填字段的重要性，在计算饱和度时必填字段会被计算两次。

4.1.4 规范性 评价上传数据的格式、内容等是否符合平台校验标准，如数据是否为空、数据类型或长度是否错误、数据编码是否符合江苏省统一制定的标准等。通过计算符合标准的数据量与总数据量的比值得出数据规范程度。

4.1.5 业务关联性 医疗卫生业务非常复杂，不同业务之间存在很深的关联，单一的评价方式难以

全面评价数据质量情况。该项通过设置多个业务关联指标，综合评价数据完整性、合理性等情况，如个人各项诊疗费用总和是否与总费用相等。

4.1.6 及时性 评价上传数据的及时性。平台要求数据每日都要进行上传，上传时间通常在凌晨，因此今天凌晨上传昨天的业务数据时及时性得分最高，如果上传的是前天的业务数据则得分降低，依次递推。

4.1.7 连续性 评价数据上传的连续性。根据每张表连续上传的天数计算得分，连续上传 1 周得满分，1 周内仅传 1 天的表得 0 分。

4.2 数据质量评价

根据评价模型的 7 个评价标准及其权重设置，计算得出每个接入机构的质量评分信息，可以了解接入机构数据质量的综合评分、各个分项评分和排名情况，见图 1。也可以了解各接入机构上传的每一张表的质量评分明细信息，见图 2。

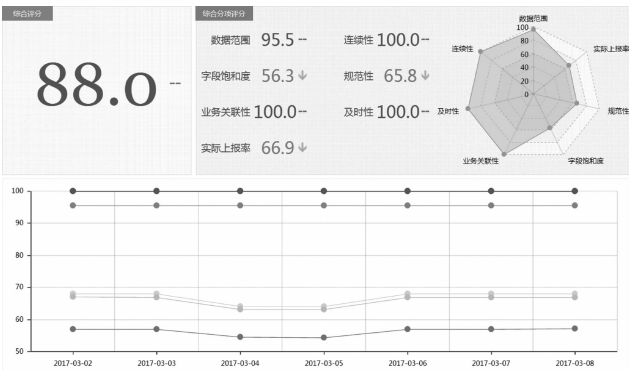


图 1 接入机构数据质量评价

医院名称	... 医院						
综合得分	88.0						
时间范围	2017-03-02 ~ 2017-03-08						
业务名称	数据范围	实际上传率	字段饱和度	规范性	业务关联性	及时性	连续性
患者基本信息数据表	是	100.0	69.1	95.3	100.0	100.0	100.0
门诊急诊挂号数据表	是	100.0	85.6	99.8	100.0	100.0	100.0
门诊急诊病历记录表	是	100.0	84.1	99.9	100.0	100.0	100.0
门诊急诊其他疾病诊断记录表	是	100.0	88.8	100.0	100.0	100.0	100.0
门诊急诊住院西药处方表	是	100.0	82.0	99.8	100.0	100.0	100.0
门诊急诊住院中药处方表	是	100.0	87.5	99.5	100.0	100.0	100.0
门诊急诊住院中药处方表	是	0.0	0.0	0.0	100.0	100.0	100.0
门诊急诊住院中药处方表	是	0.0	0.0	0.0	100.0	100.0	100.0
检验申请记录表结构	是	100.0	81.3	100.0	100.0	100.0	100.0
检验结果明细表结构	是	100.0	86.0	100.0	100.0	100.0	100.0
检查申请记录表结构	是	100.0	84.6	94.9	100.0	100.0	100.0
检查结果记录表结构	是	100.0	89.6	100.0	100.0	100.0	100.0
手术记录表结构	是	0.0	0.0	0.0	100.0	100.0	100.0
输血记录表结构	是	0.0	0.0	0.0	100.0	100.0	100.0
治疗处置记录表结构	是	100.0	89.1	100.0	100.0	100.0	100.0
抢救数据表结构	是	0.0	0.0	0.0	100.0	100.0	100.0

图 2 数据质量评分明细信息

4.3 过程追踪和反馈

4.3.1 数据上传跟踪监控 通过对每个接入机构上传数据进行全生命周期的跟踪，实时监控数据采集全流程，包括接入机构内部数据从生产、抽取、转换、上传等全过程的跟踪信息，可以及时掌握数据采集转换过程中发生的问题；对于接入机构综合评分不高的情况下，能够很快地进行问题划界与定位。图 3 展示了最近一次数据上传作业的跟踪监控信息。

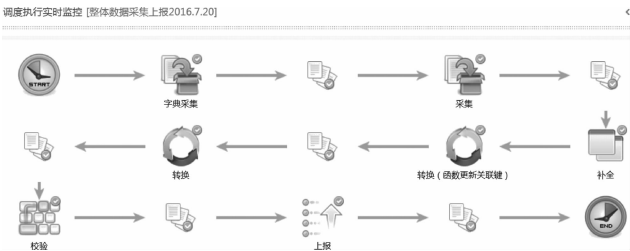


图 3 数据上传跟踪监控

4.3.2 数据质量反馈 根据平台数据标准和字典码表设置校验规则，通过数据校验模块对上传数据进行校验并生成校验报告，可以直观地掌握数据错误原因，便于快速定位数据存在的问题，能够对接入机构提供有效的改善意见，校验报告，见图 4。同时通过邮件、短信等多种通知手段，在第一时间将数据采集转换过程中发生的问题及数据校验报告、数据质量评分等情况反馈给相关接入机构，使问题及时得到改进和处理。

错误明细统计					
字段名称	规则	统计信息			
ACCOUNT_NO	SQL 必填项	正确	100%(14587)	错误	0%(0)
INVOICE_NO(发票号码)	必填项	正确	100%(14587)	错误	0%(0)
CHARGE_REFUND(收费/退费标志)	必填项	正确	100%(14587)	错误	0%(0)
OP_EM_HP_NO	枚举	正确	100%(14587)	错误	0%(0)
PAYMENT_MODE(医疗费支付方式)	必填项	正确	100%(14587)	错误	0%(0)
RCD_NAME(收费员姓名)	值域	正确	100%(14587)	错误	0%(0)
	必填项	正确	100%(14587)	错误	0%(0)

图 4 数据校验报告

5 总结

健康医疗数据具有极高价值，在大数据时代更是能够带来无法想象的收益，高质量的数据是数据共享、业务协同、大数据应用研究的基础，因此数据质量的控制和提高是人口健康信息平台建设工作中的重要环节。江苏省级平台通过新的数据质量控制体系使数据质量有了进一步的提高，但也必须注意到数据质量控制系统不能发现所有问题，部分数据质量问题必须通过人工干预才能发现，因此需从数据源头抓起，对数据从产生到应用整个生命周期的全过程加强监管与考核，使数据在业务应用中得到检验，以促进平台数据质量的提高，才能使数据质量得到根本改善。

参考文献

1 孟群. 卫生信息化建设指导意见与发展规划 (2011 - 2015) 介绍 [EB/OL]. [2016 - 12 - 08]. <http://www.docin.com/p-110825830.html>.

2 国家卫生计生委和国家中医药管理局. 加快推进人口健康信息化建设的指导意见 (国卫规划发 [2013] 32 号) [Z]. 2013.

3 李华才. 医疗健康大数据自身“健康”是关键 [J]. 中国数字医学, 2016, 11 (8): 1.

4 施天行, 王晓丽, 张寅. 浦东新区区域卫生信息数据质量现状分析 [J]. 中国卫生信息管理杂志, 2015, 12 (4): 405 - 408.

5 彭传薇, 李小华, 刘琛玺. 医院医疗数据质量现状和影响因素分析 [J]. 中国医院管理, 2005, 25 (9): 37 - 39.

6 徐汀. 数字化医院数据质量问题的分析及对策 [J]. 中国医疗设备, 2015, 30 (4): 135 - 136.

7 黄孝伦, 肖兵, 赵耀, 等. 以数据为主线探讨区域卫生信息平台建设 [J]. 医学信息学杂志, 2014, 35 (7): 7 - 11.