医疗大数据网格化清洗策略*

陈黎静 谢 桦 曹剑峰 崔 欣 刘 璐 张 玮

(上海市卫生计生信息中心 上海 200040)

〔摘要〕 以上海医疗大数据中心为依托，以质量控制之后、数据利用之前的医疗大数据为研究对象，建立数据清洗框架，给出数据可用性评估方法，根据数据特性聚类分析找出对应的清洗策略，反复推演策略的准确性、可靠性等，为医疗大数据分析利用提供有力支撑。

〔关键词〕 医疗大数据；网格化方法；数据清洗策略

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2017.02.008

Gridded Cleaning Strategy for Medical Big Data *CHEN Li-jing, XIE Hua, CAO Jian-feng, CUI Xin, LIU Lu, ZHANG Wei, Shanghai Information Center for Health, Shanghai 200040, China*

〔**Abstract**〕 Depending on Shanghai medical big data center and taking the medical big data after quality control and before data utilization as research object, the paper establishes the data cleaning frame, gives the evaluation method for data availability, finds out the corresponding cleaning strategies according to the clustering analysis of data characteristics and repeatedly deduces the accuracy, reliability of the strategy, thus providing a strong support for the analysis and utilization of medical big data.

〔**Keywords**〕 Medical big data; Grid method; Data cleaning strategy

1 引言

随着信息技术的发展，组织的管理人员进行决策分析时对数据的依赖性越来越强，于是在数据库的基础上产生了能够满足决策分析所需要的数据环境——数据仓库。但是在构建数据仓库的过程中，从异构的数据源导入的数据中存在各种质量问题，使得应用于数据仓库前端的决策支持系统产生错误的分析结果，影响信息服务的质量，因此必须对其进行数据清洗（Data Cleaning）[1]。数据清洗旨在检测数据中存在的错误和不一致等缺陷，剔除或改正缺陷，以提高数据质量[2]。有研究表明，数据清洗在数据分析的整个生命周期中占 50% ~ 80% 的比例，可见其重要性[3]。目前，国外的数据清洗研究主要集中在监测异常、重复数据，数据集成和特定领域数据清洗等方面；而国内的研究起步较晚，主要集中在特定领域内的数据清洗，如保险、金融、证券等对客户准确性要求很高的行业，研究对象往往与行业规则有很大关系[4]。

随着卫生信息化建设进程的不断加快，医疗数据的类型和规模正以前所未有的速度增长，医疗卫生领域已进入大数据时代。医疗大数据符合 IBM 提出的 4V（Volume、Value、Variety、Velocity）特征，兼具多态性、时效性、隐私性等行业特点，也

〔修回日期〕 2016-12-01

〔作者简介〕 陈黎静，博士研究生，讲师，发表论文 11 篇。

〔基金项目〕 上海市科委课题“支撑慢病创新服务和管理的大数据中心建设及数据服务模式研究与应用”（项目编号：16511102602）。

存在一切大数据所具有的数据质量问题，如空值、不一致、不完整、冗余等[5]，因此研究医疗大数据的清洗策略，使海量的信息资源更好地为医疗卫生行业的管理、诊疗、科研和教学服务，具有十分重要的意义。

近年来，上海市卫生信息化建设有序推进，通过上海市健康网工程的建设，上海市医疗大数据中心呈现 1 +17 +1 的市区两级格局：1 个公共卫生数据分中心、17 个区县数据分中心和 1 个申康数据分中心，数据分中心承载向下连接区县内各公立医疗机构或市属三级公立医疗机构、向上连接市医疗大数据中心的职能。通过市区两级共享最终汇聚了来自上海市所有公立医疗机构的数据资源。这些资源分别从不同的生产系统经过区域大数据分中心汇集后被推送至上海市大数据中心。在数据的跳转传递过程中，已经完成了前期的数据质控以及医疗大数据的初筛。本文以上海医疗大数据中心为依托，以质量控制之后、数据利用之前的医疗大数据为研究对象，建立数据清洗框架，结合数据可用性评估网格方法，根据数据特征聚类分析找出对应的清洗策略，为医疗大数据分析利用提供有力支撑。

2 数据清洗框架

2.1 概述

大数据清洗按照方式和范围不同可分为手工清洗、程序清洗和特定应用域数据清洗[2]。根据医疗大数据的行业特点，为制定准确有效的清洗策略，本研究将数据清洗框架分为数据源判别、数据可用性评估、清洗策略制定、清洗策略修正和清洗结果验证 5 个阶段，见图 1。

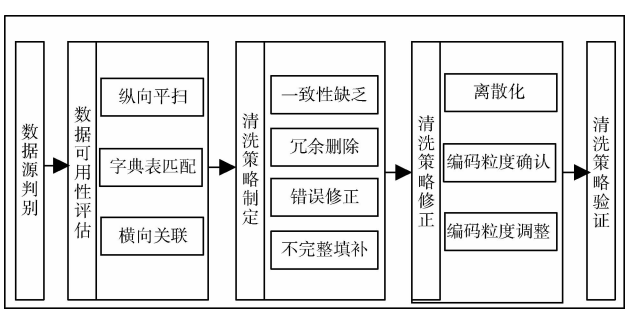

图 1 医疗大数据清洗框架

2.2 数据源判别

数据源判别需要梳理数据源状况和数据接口标准，以找到数据异同的原因，为接口升级做准备。主要内容为调查数据源生产性系统现状，包括系统种类、开发厂商、开发标准和业务规则等。上海市医疗数据的生产性系统开发厂商多达 71 家，涉及多种类。生产厂商繁杂意味着上传数据接口理解会出现偏差的可能性增大，医疗机构信息系统的开发标准差异化也增加了数据差异性表现，而且医疗机构等的业务规则差异化客观存在，繁杂的业务规则在同一接口规范下会辐射出各种类别数据。虽然部分数据可根据医疗机构上传的字典表进行转换，但总体表现为数据质量不佳。根据上述介绍可得到医疗大数据可用性不理想产生的原因主要有以下几方面：（1）医疗机构生产性系统差异性大。（2）生产性系统数据采集粒度与接口标准不一致。（3）医务人员填写数据随意性大。（4）负责上传数据的工程师上传数据主观性大。（5）接口标准定义有待完善。

2.3 网格化数据可用性评估方法

2.3.1 概述 经过前期质控规则的约束，医疗大数据的可利用状态初步显现。根据医疗大数据的行业特点，本研究给出数据可用性评估网格化方法。该方法分为纵向平扫和横向关联。纵向平扫可知晓所有数据的质量现状，可按照字段的重要程度分类评估；横向关联根据患者的一次诊疗记录，从挂号、就诊、检验检查、缴费等全流程找出数据关联性。通过独立性和关联性评估数据的可用性，为制定清洗策略提供支撑。

2.3.2 纵向平扫 可称为纯数据探查，从纯数据的角度探查可使用现状。将医疗大数据资产目录中的所有数据资产看作无内在关联的数据，从分布状况分析数据可用性，见表 1 – 表 4。

表 1 对诊疗数据集中的 1 731 937 条数据根据卡类型进行分组的统计结果

序号	卡类型	数量
1	社保卡 0	1 052 687
2	医保卡 1	254 551
3	统一自费卡 2	292 178
4	医院自费卡 3	59 146
5	其他 9	53 205
6	4	63
7	42500373300	20
8	42500377600	184
9	42500377601	52
10	42500853400	250
11	42502656400	812
12	42502658000	315
13	42502658002	2
14	42504096401	361
15	42506345300	1
16	42507230900	72
17	5	18 034
18	55740980700	4

表 2 对 1 731 937 条数据根据编码类型分组的统计结果

序号	编码类型	数量
1	–	3 828
2	–	119 022
3	nu	69 706
4	ICD – 10 01	1 367 269
5	国标 – 95 02	169 551
6	国标 – 97 03	2 561

表 3 对 686 274 条数据根据患者类型 HZLX 进行分组的统计结果

序号	患者类型	数量
1	–	131
2	–	106 274
3	本地人员 0	344 973
4	外地人员 1	169 304
5	境外人员 2	30 789
6	5	7
7	其他 9	34 796

表 4 对 1 000 000 条数据根据明细费用类别进行统计分析的结果

序号	明细费用类别	数量
1	–	5 728
2	挂号费 00	135 673
3	01	42
4	诊疗费 02	82 090
5	治疗费 03	47 441
6	04	303
7	手术材料费 05	5 257
8	检查费 06	50 026
9	化验费 07	164 976
10	摄片费 08	7 811
11	透视费 09	507
12	1	15
13	输血费 10	31
14	输氧费 11	747
15	西药费 12	225 272
16	中成药费 13	77 961
17	中草药费 14	139 524
18	其他费用 15	54 222
19	3	410
20	4	1
21	5	53
22	6	350
23	7	1 442
24	8	117
25	9	1

纵向平扫后，从可利用性角度大数据分为 3 大类：（1）无效数据。为不可用数据，如患者类型表中的“NULL”、“空”、“–”等。（2）弱可用数据。为不符合编码值域但可清洗的数据，如表 4 明细费用类别中“3、4”，再如表 1 中的“42500373300”。该类数据多由生产性系统的差异性所引起。（3）强可用数据。与完全可用数据不同，其是符合编码值域范围但准确性有待进一步验证的数据，如表 3 中的患者类型“本地人员 0”。

2.3.3 横向扫描 可称为数据关联性探查。数据关联性试图寻找单个数据与整个数据仓库的关联关系，以寻找最优清洗策略。医疗大数据关联性分为表内关联性和表间关联性。表内关联性主要考察表内数据变量间的关联。由医疗机构间的业务规则不

同导致的数据差异可设定为表间关联性。如收退费明细中医疗机构的定义差异性表现为退费分为正负值，故清洗策略可定义收退费标志和收退费明细金额为关联项。表间关联性多关注医疗大数据涉及诊疗的所有数据，包括挂号、就诊、检验检查，收费等的全流程数据。从该维度可探查数据间存在的关联性，如患者一次就诊的流水号（唯一）可贯穿全流程。两种不同的关联方式可从不同的维度给出数据清洗策略。表间关联多表现为表间外联变量与其他变量的关联性。如对 1 000 000 条数据，根据退费标识 XGBZ 关联明细项目金额 MXXMJE 的正负值进行统计分析，统计内容，见表 5。

表 5 1 000 000 条数据根据退费标识关联明细项目金额正负值的统计分析结果

退费标识别	明细项目金额 >0	明细项目金额 =0	明细项目金额 <0
收费 1	938 032	48 248	454
退费 2	11 255	668	796
0	116	83	0
3	40	0	308

2.4 数据清洗策略

2.4.1 数据分类 网格化后每个待清洗变量都在网格节点上，无效数据完全不可用，强可用数据对行业规则依赖严重，故只关注弱可用数据清洗策略的制定。具体清洗策略须针对不同类别的数据，根据编码属性将数据分为编码数据和非编码数据，见图 2。

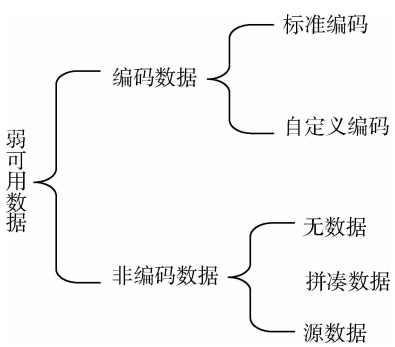

图 2 弱可用数据分类

编码数据细分为标准编码和自定义编码。标准编码为有参照标准的编码，如性别有标准编码。自定义编码为数据上传接口文档中定义了编码规则的编码，如门诊住院标志按照门诊为 1、住院为 2 的规则定义。非编码数据为没有明确编码的数据，可细分为无数据、拼凑数据和源数据（直接上传原系统数据）。无数据说明系统中没有或者有差异，上传的数据为某些极值或者默认值；拼凑数据为某些为了取得唯一性而定义的由几个变量拼凑起来的数据，这些数据拼凑后的唯一性存在，但可用性不确定；源数据多由信息系统间的差异和业务规则差异引起。

2.4.2 根据不同变量的特性确定策略 根据上述所得到的网格化结果，根据网格节点上不同变量的特性，确定该节点变量的清洗策略。归类整理后数据清洗的策略分类如下：（1）一致性缺乏的数据。该类数据可根据表间或表内关键关联字段清洗，如患者信息表中的性别、年龄等字段填报率可根据身份证号清洗，通过与人口库比对，提取年龄和性别数值进行填充。再如门诊就诊记录表中的就诊诊断编码和编码类型可通过诊断明细表中的诊断编码和编码类型补填，若两表中的数据不确定哪一个更准确，在填充数据时以完全无效数据（NULL、空、-等）为填充对象。（2）准确性不高的数据。该类未按照接口标准转换而上传的数据，因直接采集自不同的生产性系统，数据表示方式不一引起差异性较大，准确性降低。如异常提示字段（图 3）中表示异常偏高用“4”，但上传数据中表示异常偏高的数据包含“4”、“L”、“↓”、“低”等值。该类数据可分类清理，按照数据分布重新聚类，聚类后会发现机构倾向性较明显，这类数据的清洗可暂时通过脚本工具实现，或者根据辅助字段重新区分判断填充。如异常提示可根据检测指标结果与参考值对照的方法给出异常提示值。（3）符合值域范围但与其他记录对比后有明显差异的数据。该类数据多为生产性系统采集数据粒度不一所致。生产性系统采集数据粒度大于接口标准的，如诊断明细表中的诊断类型接口标准为中医或者西医，但部分医疗机构不区分中医西医，只填报中医（实际却是西医）；生产性系统采集数据粒度小于接口标准的，如门诊

住院标志除接口标准定义的门诊和住院之外还有体检数据。该两类数据可根据接口标准，若有其他选项可暂归入其他，若无，则暂不清洗（依赖接口升级），但需明确上传数据含义，可与字典表对接达到有效清洗的目的。

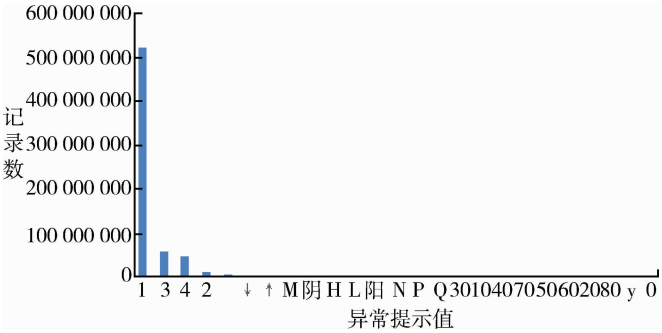

图 3 实验室检查结果表中的异常提示字段某年数据

2.5 清洗策略修正

清洗策略制定完成后需对每一条策略进行修正。本文所涉及的修正策略多集中在标准不一引起的机构倾向性明显的差异性大的数据（编码一类的数据）。由于院间的编码粒度不一，对业务规则依赖程度高，所以该类策略修正过程中需调研不同业务规则，耗费人工成本较高。如明细项目编码部分医疗机构上传的院内码，与标准接口差异大。上传院内码的数据字段可与一同上传的医院字典表针对不同机构的使用需求，将数据清洗的策略细分，有些数据需要归并的可以归并，能归并但不需要归并的可暂时先不归并，不归并可反映数据上传过程中的主要问题，如从行政管理角度是否增加细化要求。另外针对这类数据，可在数据资产目录中给出数据分布，提供给数据分析人员，既不影响分析结果，又节约清洗成本。

2.6 清洗结果验证

数据清洗是一项系统工程，在确定清洗之前，必须要对清洗策略进行反复校验，对不确定的清洗策略采用“疑罪从无”的原则。对清洗之后的结果要及时验证，对数据转换工作流和转换定义的正确性及效率进行测试和评估。可编写验证程序分步完成清洗结果验证。

3 结语

因医疗机构间业务规则定义的差异性，字典表也凸显了其在数据分析利用中的地位。字典表的治理涉及数据可用性评估及修正验证阶段，部分字典表治理需要业务部门配合。变更缓慢的字典表可由权威部门给出，灵活增加信息更新字段。如针对医疗机构代码与医疗机构名称的字典表可根据业务部门提供名单为基本字典表，医疗机构名称有变更的可在标准名称后增加相应的描述字段。

由于分析利用主题不同，数据清洗过程也有差异，利用数据清洗工具完成的清洗往往与可利用数据有较大差距，而且医疗大数据带有明显的行业领域特征，包含的行业业务规则较多，所以探索清洗工具和人工清洗策略为现阶段的一个主要任务。随着信息化进程的不断推进，数据质控规则不断修正，清洗工具不断优化，清洗策略将会在动态调整过程中不断优化、固化。

参考文献

1 RahmE, DoH H. DataCleaning: problems and current approaches [J]. IEEE Data Engineering Bulletin, 2000, 23 (4): 3 - 13.

2 蒋勋，刘喜文. 大数据环境下面向知识服务的数据清洗研究 [J]. 图书与情报，2013，(5)：16 - 21.

3 王曰芬，章志成，张蓓蓓，等. 数据清洗研究综述 [J]. 情报分析与研究，2007，(12)：50 - 56.

4 颜延，秦兴彬，樊建平，等. 医疗健康大数据研究综述 [J]. 科研信息化技术与应用，2014，5 (6)：3 - 16.

5 严英杰，盛戈皞，陈玉峰，等. 基于时间序列分析的输变电设备状态大数据清洗方法 [J]. 电力系统自动化，2015，39 (7)：138 - 144.