

面向用户需求的区域医疗数据质量综合评价研究

黄永生, 孟飞, 付晓璇, 江梦婷

(上海理工大学管理学院, 上海 200093)

摘要:当前区域医疗数据质量评价工作存在缺乏从用户需求出发进行综合评价的问题。本文在对区域医疗数据用户使用需求分析的基础上, 构建了面向用户需求的区域医疗数据质量评价框架。通过模糊综合评价对科研场景数据集质量进行实例分析, 该框架和方法可以根据用户的个性化需求, 对数据质量进行更为精确的评价。

关键词:医疗数据质量; CRITIC 法; 用户需求; 模糊综合评价

中图分类号: R195.1

文献标识码: A

DOI: 10.3969/j.issn.1006-1959.2023.10.008

文章编号: 1006-1959(2023)10-0035-05

Comprehensive Evaluation of Regional Medical Data Quality Oriented to User Needs

HUANG Yong-sheng, MENG Fei, FU Xiao-xuan, JIANG Meng-ting

(School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: At present, there is a lack of comprehensive evaluation based on user's demand in regional medical data quality evaluation. Based on the analysis of the user needs of regional medical data, this paper constructs a user-oriented regional medical data quality evaluation framework. Through fuzzy comprehensive evaluation, the quality of scientific research scene data sets is analyzed. The framework and method can evaluate the data quality more accurately according to the user's personalized needs.

Key words: Medical data quality; CRITIC method; User demand; Fuzzy comprehensive evaluation

健康医疗大数据是指在疾病防治、健康管理等过程中产生的与健康医疗相关的数据, 包括自然人从出生到死亡的完整生命过程中产生的与健康活动有关的全部数据, 涉及患者诊疗信息、病历记录或者心理健康状况等个人健康生理信息^[1,2]。在 2018 年国家卫生健康委员会发布的《国家健康医疗大数据标准、安全和服务管理办法(试行)》中, 重点强调了要充分发挥健康医疗大数据作为国家重要基础性战略资源的作用。区域健康医疗大数据在推进跨区域、跨机构、跨部门的健康医疗数据共享、政府医疗监管、临床大数据应用、科研大数据应用、公共卫生大数据应用、医疗产品与服务个性化应用等方面发挥重要作用^[3-6]。而区域医疗数据质量的相关研究主要集中在数据质量评价维度和指标体系研究方面^[7-10], 少有从用户需求角度对医疗数据质量进行评价研究^[11-14]。目前的研究工作虽然在评价体系和评价方法方面取得了一系列的研究成果, 却存在难以满足个性化用

户需求的问题。针对以上问题, 本文提出了面向用户需求的区域医疗数据质量评价框架, 并通过应用实例检测所提方法的有效性, 旨在为区域医疗数据质量评价工作提供新的研究思路和设计方法。

1 区域医疗数据质量需求

当前, 全国已建成多个省级和市级医疗数据中心, 集成了居民健康档案、电子病历、疾病控制、疾病管理、医学数字影像等业务相关数据集^[15]。经过处理的数据在医疗数据共享、政府医疗监管、临床大数据应用、科研大数据应用、公共卫生大数据应用、医疗产品与服务个性化等需求场景下发挥重要价值。但用户难以判断当前使用的数据质量是否满足要求, 另外针对特定需求场景, 不同的数据集具有不同的质量需求, 有些数据集质量较低但并不影响当前业务, 但有些数据集质量低下则不可接受^[16,17]。例如, 某些科研场景下只需要电子病历和居民健康档案业务相关数据集, 而并不关心疾病控制和疾病管理业务数据集的数据质量, 另外不同的科研任务对所使用的数据集也存在需求重要程度不同的情况。因此, 需要明确不同应用场景所涉及的数据集范围以及对数据集分配需求重要度权重。本文基于对多个项目情况的调研, 得出如表 1 所示的数据集需求矩阵。

作者简介: 黄永生(1991.3-), 男, 安徽合肥人, 硕士, 高级工程师, 主要从事医疗数据质量管理与控制研究

通讯作者: 孟飞(1982.7-), 男, 上海人, 博士, 副教授, 主要从事系统工程、人工智能方面的研究

表 1 数据集需求矩阵

数据集分类	数据集	数据共享	医疗监管	科研分析	公共卫生
居民健康档案	个人基本信息数据集	√		√	√
	职业病报告数据集	√		√	√
	...	√		√	√
电子病历	病历概要数据集	√	√	√	
	门(急)诊病历数据集	√	√	√	
	...	√	√	√	
疾病控制	艾滋病综合防治数据集		√		√
	预防接种数据集		√		√
	...		√		√
疾病管理	乙肝患者管理数据集		√		√
	肿瘤病例管理数据集		√		√
	...		√		√

2 面向用户需求的数据质量评价框架

2.1 数据质量评价流程 面向用户需求的数据质量评价流程共分为 3 个阶段,首先基于需求场景计算指标权重,其次进行单数据集数据质量综合评价,最后面向用户需求进行多数据集数据质量综合评价。各阶段主要工作说明如下。

2.1.1 指标权重计算 不同的数据应用场景需求的数据集来源和结构存在差异,从而导致在数据质量评价指标的权重设计上也需要以应用场景为单位进行设计,以保证指标权重的设置更加符合当前应用场景的需求。

2.1.2 单数据集综合评价 单数据集综合评价是以特定场景下计算好的指标权重结合模糊综合评价模型进行数据质量综合评价,最终评价结果是单数据集对评语集的综合评价向量。如果评价向量中最大元素小于要求则需进行技术处理,以消除低质量数据直到满足要求。

2.1.3 多数据集综合评价 面向特定需求场景涉及的数据集,用户可根据自身需求设定每个数据集的权重,从而形成个性化的数据集需求权重向量。然后

与单数据集综合评价形成的评价矩阵进行矩阵运算,得出最终评价结果,评价流程见图 1。

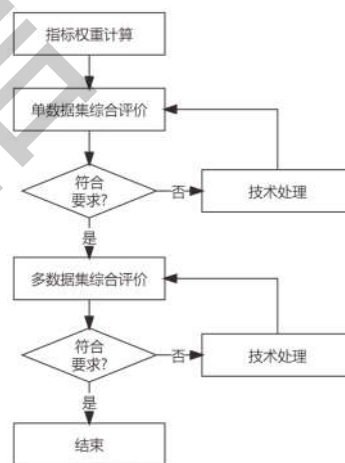


图 1 面向用户需求的数据质量评价流程

2.2 数据质量评价指标体系 本文参考《信息技术数据质量评价指标(GB/T 36344-2018)》标准对数据质量评价的要求构建了区域医疗数据质量评价指标体系,包括 4 个指标维度共 10 个指标。数据质量评价指标体系见表 2。

表 2 区域医疗数据质量评价指标体系

维度	指标	指标说明
时效性(C1)	基于时间段的正确性(D1)	基于日期范围的记录数或评率分布符合业务需求的程度
	基于时间点的及时性(D2)	基于时间戳的记录数、频率分布符合业务需求的程度
	时序性(D3)	数据集中同一实体的数据元素之间的相对时序关系
数据完整性(C2)	数据元素完整性(D4)	按照业务规则要求,数据集中应被赋值的数据元素的赋值程度
	数据记录完整性(D5)	按照业务规则要求,数据集中应被赋值的数据记录的赋值程度

表 2(续)

维度	指标	指标说明
数据准确性(C3)	数据内容正确性(D6)	数据内容是否是预期数据
	数据格式合规性(D7)	数据格式(数据类型、数值范围、数据长度、数据精度等)是否满足预期要求
	数据重复率(D8)	特定字段或数据集意外重复的度量
数据一致性(C4)	相同数据一致性(D9)	同一数据在不同位置存储或被不同用户使用数据的一致性;数据发生变化时,存储在不同位置的同一数据被同步修改
	关联数据一致性(D10)	根据一致性约束规则检查关联数据的一致性

2.3 数据质量综合评价模型 采用模糊综合评价法对特定需求场景下区域医疗数据质量进行评价^[18]。模型实现步骤包括设计评语集及隶属函数、计算指标权重、确定数据集及数据集权重、构建单数据集评价矩阵和多数数据集综合评价。

2.3.1 评语集及隶属函数 评语集指评价者对评价对象做出的所有评价结果的集合,用 V 表示。本文选取 3 个评价等级来建立评语集合,即 $V=\{\text{正常, 注意, 异常}\}$,其中 0.98 为正常、0.96 为注意、0.90 为异常。

指标的值由系统自动根据质量校验规则进行计算,以本文采用梯形分布函数,建立指标隶属度函数,值越接近 1 代表其质量满足度越高。指标对应评语集 3 种状态的隶属度函数如下:

$$\begin{aligned} A_{\text{异常}}(x) &= \begin{cases} 1, & x \leq 0.90 \\ \frac{0.96-x}{0.96-0.90}, & 0.90 < x < 0.96 \\ 0, & x \geq 0.96 \end{cases} \\ A_{\text{注意}}(x) &= \begin{cases} 0, & x \leq 0.90 \\ \frac{x-0.90}{0.96-0.90}, & 0.90 < x \leq 0.96 \\ \frac{0.98-x}{0.98-0.96}, & 0.96 < x < 0.98 \\ 0, & x \geq 0.98 \end{cases} \\ A_{\text{正常}}(x) &= \begin{cases} 0, & x \leq 0.96 \\ \frac{x-0.96}{0.98-0.96}, & 0.96 < x < 0.98 \\ 1, & x \geq 0.98 \end{cases} \end{aligned} \quad (1)$$

2.3.2 指标集及指标权重 CRITIC 分析法是一种适用于确定指标客观权重的方法,这个方法通过指标内变化的大小和指标之间的冲突来全面确定指标的客观权重^[19]。本文选择的 10 个二级指标均可收集到客观指标值,且指标之间具有一定的波动性和相关性,因此选择 CRITIC 法对指标计算权重,并通过如下几个步骤计算特定场景下指标权重。

步骤 1:基于指标之间的相关系数构建表征冲突的定量表达式来表示指标之间的冲突性。通过标准矩阵 X' 可获得每个指标的标准差 σ_i 和指标之间的相关系数 ρ_{ij} 。式中 \bar{x}_i 表示第 i 个指标的均值, $\text{cov}(X'_i, X'_j)$ 表示标准矩阵 X' 第 i 行和第 j 行的协方差。

$$\sigma_i = \sqrt{\frac{1}{m} \sum_{j=1}^m (x'_{ij} - \bar{x}_i)^2} \quad (2)$$

$$\rho_{ij} = \frac{\text{cov}(X'_i, X'_j)}{\sigma_i \sigma_j} \quad i=1, 2, L, n \quad (3)$$

步骤 2:根据 CRITIC 方法计算每个指标中包含的信息量 G_i 。式中 $\sum_{i=1}^n (1-\rho_{ij})$ 表示第 i 个指标与其他指标间冲突性的量化指标。 G_i 越大,第 i 个指标中包含的信息量就越大,该指标赋权就越大。

$$G_i = \sigma_i \sum_{i=1}^n (1-\rho_{ij}) \quad i=1, 2, L, n \quad (4)$$

步骤 3:计算客观指标 β_i 的公式如下。

$$\beta_i = \frac{G_i}{\sum_{j=1}^n G_j} \quad i=1, 2, L, n \quad (5)$$

综合步骤 1、2、3 计算得出指标权重向量为 $W=(w_1 \ w_2 \ \cdots \ w_{10})$ 。

2.3.3 数据集及数据集权重 不同需求场景涉及的数据集不同,用户可根据自身需求设定每个数据集的权重从而形成个性化的数据集需求权重向量。假设某场景下共涉及 k 个数据集,每个数据集的权重为 α_i ,则数据集权重向量 $A=(\alpha_1 \ \alpha_2 \ \cdots \ \alpha_{10})$ 。

2.3.4 单数据集综合评价 对单数据集进行模糊综合评价,分别计算它们对于评语集 V 中不同等级的隶属度 r_{ij} ,由此可得出第 i 个数据集的评价矩阵 $R_i=(r_{i1}, r_{i2}, r_{i3})$,可得当前场景下所有数据集的综合评价矩阵:

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} \end{bmatrix}$$

2.3.5 多数据集综合评价 在单数据集综合评价的基础上,结合数据集的权重向量,综合得出多数据集综合评价结果,即当前场景的所有数据集对评语集的隶属度向量 M ,其计算公式如下。

$$M = A * R = (a_1 \ a_2 \ \cdots \ a_k) \times \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} \end{bmatrix} \quad (6)$$

3 实例分析

根据上述构建的面向用户需求的数据质量评价框架,本文选取某区域医疗科研场景进行实例分析,

首先选取 60 d 的科研场景涉及的 6 个数据集(住院病案首页、入院记录、入院病程记录、检查检验记录、住院医嘱、手术治疗记录)的数据质量校验指标计算结果数据进行指标权重计算,其次选取某日待评价数据且对这 6 个数据集进行单数据集综合评价,然后对多数据集进行综合评价得出科研场景下所有数据集对评语集的隶属度矩阵,最后根据最大隶属原则判定此日数据质量评价结果。

3.1 科研场景指标权重计算 在选取 60 d 的数据中所有指标的计算结果数据后,由于指标量纲不同,需要对指标进行标准化处理,得到结果见表 3。基于特定场景下指标权重计算步骤得出科研场景下指标权重系数见表 4。

表 3 标准化后科研场景指标计算结果

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
0.7200	0.2308	0.1000	0.6219	1.0000	0.7946	0.1851	0.1200	0.8741	0.1718
0.3600	1.0000	0.9000	0.7044	1.0000	0.8200	0.4422	0.9600	0.8764	0.1348
0.2800	0.7308	0.4000	0.8541	1.0000	0.9847	0.9974	0.9200	0.8930	0.2372
...
...
0.0000	0.0385	0.2000	0.8253	1.0000	0.4517	0.4216	0.8400	0.9700	0.4494

表 4 科研场景下数据质量指标权重系数

指标项	指标变异性	指标冲突性	信息量	权重	指标项	指标变异性	指标冲突性	信息量	权重
D1	0.029	8.621	0.251	0.024	D6	0.214	8.616	1.844	0.176
D2	0.013	9.41	0.124	0.012	D7	0.189	7.823	1.482	0.141
D3	0.015	8.535	0.125	0.012	D8	0.28	9.368	2.627	0.25
D4	0.151	8.232	1.246	0.119	D9	0.127	7.888	1.005	0.096
D5	0.052	9.191	0.481	0.046	D10	0.161	8.113	1.308	0.125

3.2 科研场景模糊综合评价 通过收集到的某日指标计算结果结合隶属度函数分别对 6 个数据集进行单数据集模糊综合评价,得出单数据集对评语集隶属度矩阵如下。

$$R = \begin{bmatrix} 0.178 & 0.128 & 0.578 \\ 0.041 & 0.091 & 0.750 \\ 0.006 & 0.145 & 0.731 \\ 0.071 & 0.059 & 0.752 \\ 0.096 & 0.022 & 0.764 \\ 0.250 & 0.103 & 0.529 \end{bmatrix}$$

在单数据集综合评价的基础上,结合数据集的权重向量计算科研场景的所有数据集对评语集的隶属度向量 M 。本实例中数据集权重向量为用户指定为 $A=(0.35 \ 0.2 \ 0.15 \ 0.15 \ 0.1 \ 0.05)$,用式(6)计算科研场景所有数据集对评语集的隶属矩阵过程如下。

经过计算得到 $M=(0.104 \ 0.101 \ 0.678)$,根据以上评价等级的隶属度范围,本实例科研场景下数据质量评价结果处于“异常”水平。

$$M=A \times R = \begin{pmatrix} 0.35 & 0.2 & 0.15 & 0.15 & 0.1 & 0.05 \end{pmatrix} \times \begin{pmatrix} 0.178 & 0.128 & 0.578 \\ 0.041 & 0.091 & 0.750 \\ 0.006 & 0.145 & 0.731 \\ 0.071 & 0.059 & 0.752 \\ 0.096 & 0.022 & 0.764 \\ 0.250 & 0.103 & 0.529 \end{pmatrix}$$

4 总结

本文将用户需求贯穿于区域医疗数据质量评价的全过程,明确了面向用户需求的数据质量评价框架。首先,基于用户的数据需求设计了数据质量评价流程,构建了数据质量评价指标体系。其次,建立了数据质量综合评价模型,利用 CRITIC 分析法确定了指标权重,同时确立了单数据集和多数据集综合评价矩阵。最后,通过某区域科研场景对数据质量进行了综合评价,其评价结果总体上为“异常”水平。总体来看,本文构建的采用面向用户需求的区域医疗数据质量评价框架既能够根据用户需求灵活地进行数据质量评价,也能够及时根据用户需求及反馈信息中隐含的规律动态完善评价指标权重。根据这一研究结论,提出如下建议:①在区域医疗数据评价和管理过程中应多关注和了解用户的真实需求,结合用户需求来决定对哪些数据集进行评价以及动态调整评价指标的权重等;②在对数据质量进行评价时,可以考虑让用户参与进来,根据自身需求对数据集的重要程度进行辨识和赋权,使评价结果更贴近用户需求。未来研究可对数据质量评估框架的测度及实现方法进行细化,同时探索纳入用户对数据质量定性评价指标,如用户的质量感知性评价,另外需将本文所构建的面向用户需求的区域医疗数据质量评价框架下沉到实践应用中去,在实践中不断完善框架的适用性。

参考文献:

[1]叶清,刘迅,周晓梅,等.健康医疗大数据应用存在的问题及对策探讨[J].中国医院管理,2022,42(1):83-85.
[2]郑忆,张书铭,赵梦莹.基于典型应用场景的健康医疗大数据安全保障体系研究[J].电子技术与软件工程,2022,220(2):9-12.
[3]王立梅.健康医疗大数据的积极利用主义[J].浙江工商大学学报,2020,162(3):32-40.
[4]El Emam K,Paton D,Dankar F,et al.De-identifying a public use microdata file from the Canadian national discharge abstract database[J].BMC Med Inform Decis Mak,2011,11:53.

[5]Chan KS,Fowles JB,Weiner JP.Review: electronic health records and the reliability and validity of quality measures: a review of the literature [J].Med Care Res Rev,2010,67 (5):503-527.
[6]Reimer AP,Milnovich A,Madigan EA.Data quality assessment framework to assess electronic medical record data for use in research[J].Int J Med Inform,2016,90:40-47.
[7]Toivonen M.Big data quality challenges in the context of business analytics[D].Helsinki:University of Helsinki,2015.
[8]Johnson SG,Speedie S,Simon G,et al.A Data Quality Ontology for the Secondary Use of EHR Data [J].AMIA Annu Symp Proc,2015,2015:1937-1946.
[9]Johnson SG,Speedie S,Simon G,et al.Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data[J].Appl Clin Inform,2016,7(1):69-88.
[10]高亮.数据治理:让数据质量更好[J].中国教育网络,2014,115 (12):64-66.
[11]舒婷,刘海一,赵韡.电子病历系统功能应用水平分级评价标准修订思路探讨[J].中华医院管理杂志,2018,34(3):198-200.
[12]王雯璟,沈绍武.电子病历共享文档规范应用研究[J].湖北中医杂志,2014,36(3):78-80.
[13]李健,王明月,许路明,等.基于用户感知价值的医疗信息服务评价体系构建[J].数据分析与知识发现,2019,3(2):118-126.
[14]王巍,宿明.基于用户感知价值的医疗信息服务指标体系的构建[J].现代情报,2017,37(2):19-24.
[15]梁佩丽,肖继红.住院病案首页数据辅助医院医疗质量评价模型的应用效果[J].中国病案,2022,23(6):27-31.
[16]张世红,李磊,史森.区域健康医疗大数据中心体制机制研究[J].医学信息学杂志,2020,41(5):43-48.
[17]莫祖英,邝苗苗.基于用户视角的政府开放数据质量评价模型及实证研究[J].大学图书馆学报,2020,38(4):84-89.
[18]郭爽.基于模糊综合评价法的港口智慧化建设需求评价[J].物流技术,2022,41(4):89-94,125.
[19]赵洪山,李静璇,米增强,等.基于 CRITIC 和改进 Grey-TOPSIS 的电能质量分级评估方法[J].电力系统保护与控制,2022,50(3):1-8.

收稿日期:2022-07-11;修回日期:2022-08-25

编辑/杜帆