区域卫生数据用于临床疗效分析的可用性研究

叶琪[1]，赵亮[1]，阮彤[1]，冯东雷[2]，高炬[3]，刘珉[3]
1. 华东理工大学，上海 200237；2. 万达信息股份有限公司，上海 200040；
3. 上海中医药大学附属曙光医院，上海 201203

摘要
区域卫生数据来源于不同的医院，经历多轮采集、转换、整合过程，数据可能存在各种质量问题。提出了一套“数据可用性”的评估流程，从临床科研需求出发，设计评估需求模板，采集评估需求，进而获得评估度量。通过面向心力衰竭疗效分析的需求实例，获得了与数据的完整性和一致性相关的10个度量指标，并对某省级区域平台的数据进行了可用性评估。结果发现，与临床科研相关的数据的完整性和一致性都仍有待提高。

关键词
区域卫生数据；数据质量；数据可用性；疗效分析

中图分类号：TP391　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018026

Usability research of regional health data for clinical efficacy analysis

YE Qi[1], ZHAO Liang[1], RUAN Tong[1], FENG DongLei[2], GAO Ju[3], LIU Min[3]
1. East China University of Science and Technology, Shanghai 200237, China
2. Wonders Information Co., Ltd., Shanghai 200040, China
3. Shanghai Shuguang Hospital, Shanghai 200203, China

Abstract
The regional electronic health records data are initially collected from several different hospitals and then undergo several rounds of data collection, transformation and integration, so there may exist a variety of quality problems. A new evaluation process for data usability was presented. The evaluation metrics were obtained by firstly gathering research requirements of clinicians, and then designing of requirement templates in the process. An example of obtaining evaluation metrics based on the requirement analysis with heart failure was also given. By using the process, ten metrics related to the data completeness and consistency were obtained and then the data of the provincial regional platform were evaluated for the usability. Finally, that the completeness and the consistency of data related to clinical research needs to be improved was shown.

Key words
regional health data, data quality, data usability, efficacy analysis

1 引言

在过去的十余年中，为促进医疗信息的互联互通与信息共享，加强卫生监管监督，各省市大多建立了区域卫生信息平台。经过若干年的积累，这些平台已经收集和存储了区域内医院的大量电子病历（electronic medical record，EMR）数据，这些数据不但记录了病人的基础信息，也记录了医生的诊断信息、患者的用药等治疗信息。将这些数据用于临床科研，对激活区域卫生数据有重要的作用。临床科研人员可以通过对海量临床电子病历数据的分析发现病人治疗的效果、病情的变化情况，从而促进新治疗方案的提出和药物的研发。近几年各国政府正在积极推动使用电子病历进行临床研究，例如，美国国立卫生研究院推动的战略健康IT高级研究项目（Strategic Health IT Advanced Research Projects，SHARP）[1]等。而在疗效分析、转归分析和临床工作流分析等临床科研中，基于电子病历的研究也受到广泛的关注[2-4]。

由于区域卫生平台上的数据来源于医院电子病历系统，或医院多个信息系统融合而成的临床数据中心（clinical data repository，CDR），这些系统在医院发展过程中经历了多次版本变化，而在不同系统之间，数据的统一表示、关联和集成都存在各种问题。区域卫生平台数据又来源于不同医院，异构性更强。同时，医生录入电子病历主要记录医疗过程，记录过程中数据本身会有缺失，而如何记录和表示这个过程，也缺乏语义规范。

另外，无论是区域平台的构造，还是电子病历本身，都并不直接面向科学研究。从管理收费角度来看质量良好的数据可能无法满足临床科研的需求。因此，为了掌握区域平台数据能否或在多大程度上支撑某一类临床科研，就需要在做科研之前，对数据进行质量评估。评估结果可以用于设计与细化实验分析过程，也可以用于未来改造医学系统，提升电子病历数据质量。

目前，电子病历的研究工作集中在提出电子病历数据的度量指标以及利用指标发现数据集合上的质量问题[5]。例如，Weiskopf N G等人[6]提出了Breadth、Documentation、Density、Predictive等多个面向数据完整性的指标，并在医院数据上进行了验证。参考文献[7]提出电子病历数据中存在的错误和不准确主要包括数据实体错误、数据字段缺失、整个记录缺失以及数据抽取和变换过程引发的错误。参考文献[8]提出了面向电子病历的3个数据质量维度：一致性、完整性和真实性，从外部规定和内部特征两个方面给出了质量度量指标，并给出了每个度量指标的实例，但这些度量指标没有应用在具体的数据集中。参考文献[9]不仅从完整性、准确性和及时性3个方面定义了13个评估指标，还利用这13个指标对支持iSantéé系统的95家机构的数据进行了定量评估，发现了各机构的数据质量问题，并给出机构的数据质量排名。

在上述研究中，使用哪些质量指标评估数据完全是根据评估者的经验。根据数据质量公认的定义[10]，质量是“数据适合使用的程度（fit to use）”，数据使用的场景是数据质量的核心，数据质量评估的结果应反映数据与使用的适合程度。为此，本文提出了一套基于“数据可用性”的通用评估流程，该流程包括评估需求模板的设计、评估需求的采集以及从评估需求中得到的评估度量。根据此评估流程，可以获得一系列评估度量。

基于该方法，针对某个省级区域平台的心力衰竭病人数据的临床疗效分析需求，做了数据质量可用性评估。

2 相关工作

Wang R Y[11]最早提出了一种基于过程管理的数据质量管理（total data quality management，TDQM）方法，引入了工程管理中质量管理的方法，全面管理数据的质量。该方法分为4个步骤：定义、评估、分析、提高。该方法面向数据库用户建立一套信息质量度量体系，评估数据错误率、数据丢失率，并记录一致性和最近数据情况等。随后Lee Y W等人[12]提出了信息质量评估方法（methodology for information quality assessment，AIMQ），帮助企业全面了解信息质量状况，并能随时监视信息质量的改进过程。AIMQ将信息按信息产品质量和信息服务质量、符合规范和满足用户期望情况分为4类：健全信息、有用信息、可信赖信息、可用信息。通过专家调研得到每类信息的质量维度，构建产品服务性能/信息质量（product service performance / information quality，PSP/IQ）模型。TDQM和AIMQ从不同角度分别建立了信息质量度量体系，但这两种方法均没有涉及评估需求的产生过程，且建立的指标体系与应用场景无关。

考虑具体的应用场景，Batini C等人[13]提出了一种综合数据质量管理方法——CDQM（complete data quality methodology）。CDQM关注数据与组织之间的密切关系，认为数据质量评估和改进与业务流程和组织密切相关。该方法将评估分为3个阶段：数据状态的重构、评估和选择最优的改进流程。数据状态的重构阶段重新建立组织机构、流程、服务和数据之间的关系；评估阶段根据与提供服务有关的主要问题定义新的质量维度，评估数据流和数据库的数据质量，并确定质量问题所在的位置。该方法能快速发现产生质量问题的数据所在位置，为提升数据质量提供了依据。此外，Woodall P等人[14]提出了一种数据质量评估的混合方法，目的是省略不必要的评估步骤，节约数据质量评估的成本、时间和资源。通过抽取现有评估技术中的活动，对不同的活动进行分类，然后明确评估目标，根据需求选择所需的活动，省略不必要的活动或者将这些活动并行运行，组合成新的评估技术。上述研究的评估目标主要是提升企业的数据质量，要求数据符合管理需求。但在实际应用中，数据集除了适用于管理，还有许多其他用途。

在医疗领域中，参考文献[15]针对衣原体病和淋病患者的电子报告（ePR）进行数据质量评估。评估过程分为5个步骤：验证报告文件格式，确认所需字段都存在；分析ePR记录是否都可报告；分析ePR记录是否都能正确显示；平行验证历史数据中患者地址、治疗日期、治疗信息、种族、民族的完整性，测试ePR每日数据记录的完整性和及时性；重新评估过程，对实验室报告记录进行再次验证，并将结果作为金标准。参考文献[16]对电子病历的数据用于产科研究的可靠性进行了评估。研究者首先从电子病历报告中提取数据，然后根据经验定义了评估指标：分娩方法、诱导产程、宫颈成熟、顶点呈现和产后出血等，对评估指标的数据一致性进行了评估。由于在医疗领域中每种应用场景的质量要求各不相同，评估的度量体系也各不相同，因此，需要定义一套可操作的通用评估流程，分析数据集在不同的应用场景中是否适用。

3 评估方法

本文提出的数据可用性评估方法如**图1**所示。首先，通过评估需求模板收集评估需求，汇聚需求得到评估字段；其次，利用数据集成和数据筛选过程确定评估需求；再次，建立评估映射关系并定义评估度量，定义评估度量后可返回更新评估映射；最后，利用错误自动检测技术定量地评估数据质量，并根据评估结果分析数据的可用性。

整个过程可分为以下6个步骤：

步骤1 收集评估需求；

步骤2 确定评估数据；

步骤3 建立评估映射；

步骤4 定义评估度量；

步骤5 评估数据质量；

步骤6 分析评估结果。

整个过程主要由3类角色参与，第一类角色是领域专家（例如临床医生），其负责步骤1，参与步骤3和步骤5。第二类角色是数据管理者以及负责系统构建与数据集成的信息技术（information technology，IT）工程师，其负责步骤2和步骤3，同时参与步骤5。第三类角色是构造与执行质量度量的数据质量工程师，其负责步骤4和步骤5。对于步骤6，需要3类角色共同参与，解释数据结果。

3.1 收集评估需求

为了采集评估需求的具体内容，本文定义了一个评估需求模板，见**表1**。质量评估需求模板包括：需求编号、评估

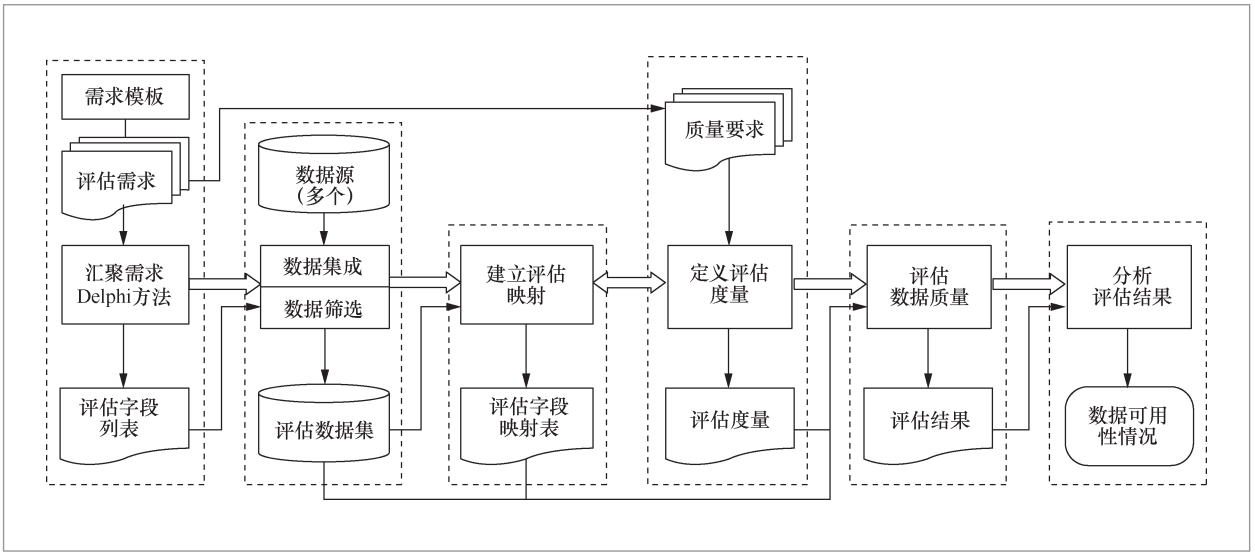

图1 数据可用性评估方法示意

表1 评估需求模板

需求编号	评估字段类别	评估字段	质量要求	字段重要程度	语义相关字段	需求来源
数字	诊断信息	诊断名称、诊断代码等	诊断代码必须是ICD10的编码	重要/一般	病人主诉	诊断和治疗指南/临床科研人员

字段类别、评估字段、质量要求、字段重要程度、语义相关字段与需求来源。评估需求的核心是评估字段和质量要求。例如，评估字段是临床诊疗中录入或使用的数据字段；质量要求是基于业务规则或医疗信息标准，对数据字段质量相关特性的描述。字段类别描述了评估字段属于哪一类信息，需求来源是提供评估需求的人或参考资料，语义相关字段描述了与评估字段发挥类似作用的字段信息。

通过Delphi过程确定最终评估需求。Delphi方法[17]又被称为专家调查法，它利用多轮专家咨询调查，能快速汇集不同的意见，具有针对性强、科学性强等特点，被用于关键因素分析[18]、诊疗方案的有效性调查[19,20]等领域。在本文中，首先，选择第一类和第二类角色作为领域专家；其次，临床科研人员填写评估需求模板中的评估内容，数据质量工程师根据《中国心力衰竭诊断和治疗指南2014》[21]和文献资料补充评估内容，数据管理人员和IT工程师补充需求模板中的质量要求，最终根据评估需求模板进行多轮调查，确定对临床疗效分析研究重要的评估内容。

面向心力衰竭的疗效分析研究的需求，基于Delphi过程，最终得到的评估字段包括年龄、性别、诊断编码、疾病名称、体征信息（血压、心率）、检验信息（血脂、脑钠尿肽（brain natriuretic peptide，BNP）等）、影像检查（心电图、心脏彩超）和疗效信息。

3.2 确定评估数据

评估数据来源于某个省级区域平台，该平台对该省内的多家医院内部的医院信息系统（hospital information system，HIS）、临床信息系统（clinical information system，CIS）、实验室信息管理系统（laboratory information management system，LIS）和放射信息系统（radiology information system，RIS）等原始业务数据进行整合，然后根据统一的接口规范进行整合处理。整合过程需纠正部分数据（例如医院代码等），将整合后的数据上传到中心数据库。

本文关注了区域平台的心力衰竭病人数据，选择了2013—2016年在区域平台覆盖的医院门诊就诊或住院的患者，经过筛选最终得到大约380万条记录，其中住院记录122 177条，门诊就诊记录3 673 222条，病人大约26万人。

3.3 建立评估映射

建立步骤1中评估字段与步骤2中评估数据集之间的映射关系。针对**表1**的需求字段，根据区域平台的数据交换接口规范，说明评估字段的数据来源，如果评估字段存在于多个数据源中，则选定评估字段来自哪个数据源；说明评估字段是结构化数据类型、文本类型还是影像类型。

3.4 定义评估度量

针对心力衰竭疗效分析需求，根据步骤1中医生对评估需求模板的反馈，给出心力衰竭疗效分析的评估度量指标体系。根据前面得到的评估需求字段删除评估需求中无数据来源的评估字段，选择、组合或自定义评估指标，最终得到患者的基本信息（出生日期、性别）、体征信息（血压、心律）、诊断信息（诊断编码、诊断名称）、检验信息（血脂、BNP）和疗效信息（治疗效果、死亡信

息）共计10个指标。

上述10个评估指标利用参考文献[22]提出的金标准、数据元素存在性、数据元素一致性、数据源一致性等评估方法，根据**表1**中得到的质量要求（评估规则），从数据的完整性和一致性两个方面进行评估。其中，完整性评估判断评估字段值是否为空、是否为无意义的系统默认值，评估度量见**表2**。

一致性评估判断评估字段是否遵循了统一的规范、是否保持了统一的格式以及不同数据来源的相同内容是否一致，评估度量规则见**表3**。

表2和**表3**中定义的度量规则可以是复合规则（复合规则由多个基本规则组成）。

3.5 评估数据质量

本文采用基于错误自动检测技术[23,24]的比率法评估数据质量。根据**表2**和**表3**定义的完整性和一致性的评估度量规则进行错误检测，执行查询语句，发现违反规则的所有记录，迅速发现数据集中的错误，计算出符合度量规则的记录数与总记录数

表 2 评估指标的完整性度量规则

评估指标	度量规则
出生日期	出生日期不为空
性别	性别字段编码不为空，应为“1”或“2”
血压	入院时主要症状及体征信息包含“血压%”或“BP%”
心律	入院时主要症状及体征信息包含“心律%”或“HR%”
诊断编码	诊断编码不为空且不等于“自定义”或“-”
诊断名称	诊断说明不为空且不等于“null”
血脂	同一流水号的收费项目中包含血脂信息，且实验室检查报告中存在一条报告类别名称为血脂信息的记录
BNP	同一流水号的收费项目中包含“BNP”且实验室检查报告中存在一条报告类别名称为“BNP”的记录
治疗效果	治疗效果不为空
死亡信息	死亡原因不为空且不等于“0”，或者死亡时间不为空且不等于“1900”

表 3 评估指标的一致性度量规则

评估指标	度量规则
出生日期	患者基本信息与病案首页中出生日期的数据格式统一； 同一患者在患者基本信息与病案首页中出生日期相同
性别	同一患者在患者基本信息表与病案首页中性别相同； 同一患者在患者基本信息表与实验室检查报告中性别相同； 同一患者在病案首页（旧）与病案首页（新）中性别相同； 同一住院患者在患者基本信息表、病案首页、实验室检查报告、医学影像检查报告、出院小结表中性别均相同； 同一门诊患者在患者信息表、实验室检验报告、医学影像检查报告中性别相同
诊断编码	属于GB/T 14396疾病编码
诊断名称	属于GB/T 14396疾病名称

之间的比率。评估算法描述如下。

输入：评估数据集S_0，评估指标i，评估指标i的度量规则表$\{R_{i1}\cdots R_{in}\}$。

输出：评估结果V_i。

方法：

for each R_{ij} in $\{R_{i1}\cdots R_{in}\}$

{ 生成R_{ij}的SQL查询语句SQL_{ij};

对S_{j-1}执行SQL_{ij}，得到违反规则的错误数据集S_j' ;

输出错误数据集S_j' ;

$S_j=S_{j-1}-S_j$' ;}

next

V_i=Count(S_n)/Count(S_0)

4 评估结果

从总体上看，本次评估的某区域平台的心力衰竭患者数据在疗效分析研究中具有一定的可用性。一方面，区域卫生数据为临床研究提供了丰富的样本数。本次评估的心力衰竭的患者总数达到26万，记录数达380万。另一方面，数据整体完整性水平较高，评估结果如**图2**所示。患者基本信息中的出生日期和性别完整性分别达到75.18%和99.82%。诊断信息中的诊断编码、诊断名称完整比例分别为80.18%和82.31%。治疗结果的完整性高达99.44%。但是，现有区域数据质量存在的一些问题制约了数据的使用。具体包括以下几个方面。

（1）部分数据的完整性存在问题

血压和心率数据缺失较多，血压和心率数据完整性分别为16.34%和4.96%，出现这一情况是由于这两个指标不直接影响医院的管理，区域卫生平台并未强制要求提交相关信息，部分医院的血压和心率数据没有集成到区域卫生平台中。此外，死亡信息的完整性也较低（比例为7.92%）。主要原因是受传统风俗观点的影响，很多重症患者临终前会选择回家，在医院死亡的患者比例较低。

（2）检验项目的分类和名称的统一会提高数据的可用性

检验项目中血脂和BNP数据完整性不能完全反映数据的真实情况。见**表4**，血脂和BNP检验收费记录占总检验记录的比例分别为11.04%和2.45%。明显可以看出，无论是有这两项检验收费的记录数还是包含这些检查项目的报告数都较少，与患者的实际就诊情况不符。

出现上述情况的原因是各家医院检验

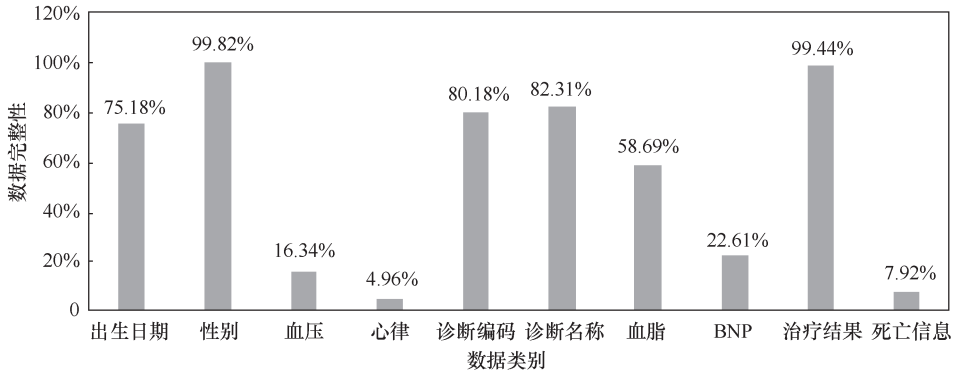

图 2 数据整体的完整性

的项目分类不尽相同，项目名称也是由各家医院自行规定的，由于分类和名称不统一，所以无法得到准确的统计数据。例如，血脂的检验项目包括多种，不同医院血脂包含的指标个数和指标名称各不相同。血脂检验项目可以包括总胆固醇、甘油三酯、高密度脂蛋白、低密度脂蛋白4项检验指标，也可以包括总胆固醇、甘油三酯、高密度脂蛋白胆固醇、低密度脂蛋白胆固醇、载脂蛋白AI、载脂蛋白B 6项指标。每项指标名称也不统一，例如高密度脂蛋白可记为“高密度脂蛋白胆固醇”“HDL”“HDL-C”等。由于没有统一的标准，在提交的收费记录和检验信息中，有的医院使用血脂这一项目名称，有的医院使用具体的检验指标项目，检验名称多达10余种，造成数据分析的大量误差。检验项目数据完整性情况见**表4**。

（3）诊断术语的不统一影响数据的可用性

虽然区域卫生平台要求医院提交的诊断数据必须遵守国家标准GB/T 14396（可看作中国版ICD-10），但使用标准中的诊断编码和诊断名称进行检索时，仍然无法获得完整的诊断数据。将诊断编码、诊断名称与GB/T 14396标准比较，发现与标准不相符的编码和名称比例分别为17.38%和81.81%。出现这种情况的主要原因是现有ICD-10的编码内容不够丰富，不能满足临床和管理的需要。现有编码内容无法精准描述疾病的精确部位、是首次确诊还是复发、是否做了手术等疾病状况。虽然部分医院通过扩展院内编码或者扩展疾病名称的方式满足了临床需求（此时院内信息系统上传时需要与标准GB/T 14396进行转换），但是部分医院将院内编码直接上传造成了部分编码与标准不一致。

（4）虽然平台已制定数据交换接口规范，但部分数据提交时未遵守规范

接口规范规定性别编码“1”代表男性，“2”代表女性，实际数据中同一患者性别编码不相同的比例为42.67%。规范规定出生日期的格式为“YYYYMMDD”，但实际数据表示各不相同。造成这一现象的原因是数据的产生部门及录入途径不同，且各业务子系统的建设时期不同，制定的内部数据规范也各不相同，而且医院进行数据整合时也没有将内部数据按照规范要求进行格式转换或数据映射。

为了提高数据的可用性，需要从多方面入手提高数据的质量。可以考虑从以下几方面进行改进。

首先，区域卫生平台还需要集成更多的业务数据。血压和心率数据是心血管疾病的重要指标之一，现有平台中这两个指标的数据缺失较多，需要部分医院将血压和心率数据集成到区域卫生平台。此外，影像数据（心电图和心脏彩超）直接影响着心血管疾病的预后（也是疗效评估的依据），对疗效研究发挥着重要作用。而现有系统中缺乏此类检查数据，区域数据平台如果能集成这些数据，将会增强疗效分析研究数据的可用性。

表4 检验项目数据完整性情况

检验指标	检验报告		收费记录		检验记录总数/个
	报告数/收费记录	报告数/个	收费记录/检验记录总数	记录数/个	
血脂	58.69%	412 744	11.04%	703 306	6 369 697
BNP	22.61%	35 343	2.45%	156 308	6 369 697

其次，建立或使用统一规范的元数据。国际疾病分类（international classification of diseases，ICD）编码与名称一致性比较低，而检验项名称也缺乏统一规范，需要大量的数据后处理工作。未来需在标准的疾病细分层面的基础上使用规范的ICD编码，全面使用卫生和计划生育委员会正在积极推广的新版标准《GB/T 14396-2016 疾病分类与代码》。引入一套完整的检验项名称和编码规则，改善相关数据的质量。此外，补充现有临床术语或者编码与标准术语或规范之间的映射表，实现数据的匹配和映射。

最后，建立数据质量管理的闭环机制。各医院严格遵守数据上传规范，减少数据的缺失和格式的不统一，同时区域平台定期给出数据质量情况分析，敦促各医院不断提高上传数据的质量。

5 结束语

本文提出了一种基于实用的标准化数据质量评估框架，通过引入预定义的评估需求模板收集评估需求，并利用Delphi过程快速汇聚各方共识，最终形成广泛认可的评估需求，然后根据评估需求采集评估数据，建立映射关系，再定义评估度量，完成评估任务。该评估方法回答了数据集对不同应用场景是否可用的问题，虽然仅评估了从某区域数据平台采集的数据集对心力衰竭临床疗效研究是否可用，但是该评估方法不仅适用于评估临床疗效这一场景，还适用于其他的应用场景。

未来重点考虑利用SNOMED CT、LONIC编码作为临床术语表示的标准，建立现有数据与标准的映射表，解决术语不统一的问题。

参考文献:

[1] REA S, PATHAK J, SAVOVA G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data[J]. Journal of Biomedical Informatics, 2012, 45(4): 763-771.

[2] BROWN J S, HOLMES J H, SHAH K, et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care[J]. Medical Care, 2010, 48(6): 45-51.

[3] BAYLEY K B, BELNAP T, SAVITZ L, et al. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied[J]. Medical Care, 2013, 51(8): 80-86.

[4] MICHELLE R H, SARAH R B, ISAAC H G, et al. Secondary use of electronic health record data for clinical workflow analysis[J]. Journal of the American Medical Informatics Association, 2018, 52(1): 40-46.

[5] REIMER A P, MILINOVICH A, ELIZABETH A M. Data quality assessment framework to assess electronic medical record data for use in research[J]. International Journal of Medical Informatics, 2016, 90(6): 40-47.

[6] WEISKOPF N G, HRIPCSAK G, SWAMINATHAN S, et al. Defining and measuring completeness of electronic health records for secondary use[J]. Journal of Biomedical Informatics, 2013, 46(5): 830-836.

[7] SUKUMAR S R, NATARAJAN R, FERRELL R K. Quality of big data in health care[J]. International Journal of Health Care Quality Assurance, 2015, 28(6): 212-218.

[8] 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]. EGEMS(Wash DC), 2016, 4(1): 1244-1261.

[9] PUTTKAMMER N, BASEMAN J G, DEVINE E B, et al. An assessment of data quality in a multi-site electronic medical record system in Haiti[J]. International Journal of Medical Informatics, 2016, 86(2): 104-116.

[10] JURAN J M. Juran on planning for quality[M]. New York: Free Press, 1988: 14-23.

[11] WANG R Y. A product perspective on total data quality management[J]. Communications of the ACM, 1998, 41(2): 58-65.

[12] LEE Y W, STRONG D M, KAHN B K, et al. AIMQ: a methodology for information quality assessment[J]. Information & Management, 2002, 40(3): 133-146.

[13] BATINI C, SCANNAPIECO M. Data and information quality[M]. Heidelberg: Springer Press, 2016: 353-402.

[14] WOODALL P, BOREK A, PARLIKAD A K. Data quality assessment: the hybrid approach[J]. Information& Management, 2013, 50(7): 369-382.

[15] TSENG M, RAKETICH N, SIMMONS C. Evaluation of chlamydia and gonnorhea electronic provider reports data quality[J]. Journal of the Public Health Information, 2018, 9(1): 337-344.

[16] ALTMAN M R, COLORAFI K, DARATHA K B. The reliability of electronic health record data used for obstetrical research[J]. Applied Clinical Informatics, 2018, 9(1): 156-162.

[17] DALKEY N C, ROURKE D L. Experimental assessment of Delphi procedures with group value judgements [M]. Santa Monica: Advanced Research Projects Agency, 1971: 1-58.

[18] MCGINN C A, GAGNON M P, SHAW N, et al. Users' perspectives of key factors to implementing electronic health records in Canada: a Delphi study[J]. BMC Medical Informatics and Decision Making, 2012, 12(11): 1-13.

[19] 王磊, 张俭, 郭力恒, 等. 基于改良德尔菲法对急性心肌梗死中医诊疗方案的调查分析[J]. 中西医结合心脑血管病杂志, 2010, 8(10): 1165-1167.
WANG L, ZHANG J, GUO L H, et al. Investigation and analysis of traditional Chinese medicine diagnosis and treatment plan for acute myocardial infarction based on modified Delphi method[J]. Chinese Journal of Integrative Medicine on Cardio-/Cerebrovascular Disease, 2010, 8(10): 1165-1167.

[20] KERRY N L, KATY A C, SARA T B, et al. Development of a core outcome set for clinical effectiveness trials in esophageal cancer resection surgery[J]. Annals of Surgery, 2018, 267(4): 700-710.

[21] 中华医学会心血管病学分会. 中国心力衰竭诊断和治疗指南2014[J]. 中华心血管病杂志, 2014, 42(2): 3-10.
Chinese Society of Cardiology. Chinese heart failure diagnosis and treatment guide 2014 [J]. Chinese Journal of Cardiology, 2014, 42(2): 3-10.

[22] WEISKOPF N G , WENG C H. 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(2): 144-151.

[23] CHEN W, FAN W, MA S. Analyses and validation of conditional dependencies with built-in predicates[C]// Database and Expert Systems Applications, August 31-September 4, 2009, Linz, Austria. Heidelberg: Springer Press, 2009: 576-591.

[24] KONTOKOSTAS D, WESTPHAL P, AUER S, et al. Test-driven evaluation of linked data quality[C]// The 23rd International Conference on World Wide Web, April 7-11, 2014, Seoul, Korea. New York: ACM Press, 2014: 747-758.

作者简介

叶琪(1976-),女,博士,华东理工大学讲师,主要研究方向为大数据、数据质量等。

赵亮(1993-),男,华东理工大学硕士生,主要研究方向为大数据、数据质量等。

阮彤(1973-),女,博士,华东理工大学计算机技术研究所教授、所长,主要研究方向为数据质量评估、文本抽取、知识图谱等。

冯东雷(1972-),男,博士,万达信息股份有限公司教授级高级工程师,主要研究方向为健康医疗大数据+人工智能、健康医疗+互联网、卫生信息标准化等。

高炬(1966-),男,上海中医药大学附属曙光医院副院长、主任医师,主要研究方向为医院行政管理、中西医结合肝胆病研究。

刘珉(1964-),男,上海中医药大学附属曙光医院信息科主任,主要研究方向为医疗信息管理。

收稿日期: 2018-03-20

基金项目: 上海市科学技术委员会“科技创新行动计划”高新技术领域基金资助项目(No.16511101000);上海市卫生和计划生育委员会中医药创新基金资助项目(No.ZYKC201601013)

Foundation Items: “Action Plan for Innovation on Science and Technology” Projects of Shanghai (No.16511101000), Traditional Medicine Innovation Project of National Health and Family Planning Commission of Shanghai (No.ZYKC201601013)