·综述·

基于医疗保险数据的数据库准确性验证方法学进展

冯菁楠　王胜锋　詹思延
北京大学公共卫生学院流行病与卫生统计学系 100191
通信作者：詹思延，Email：siyan-zhan@bjmu.edu.cn

【摘要】 医疗保险数据库蕴藏着丰富的信息，是研究人群疾病特征、疾病负担、提供管理政策制定依据的重要来源。在医保数据库中，通常利用疾病编码和名称构建算法来识别患者，因此，数据库准确性的验证对判断算法是否正确识别所研究疾病或某种暴露因素的人群十分重要。本文介绍国外传统的病历审查方法，并结合机器学习、自然语言处理及数据库链接等新兴辅助验证技术，探讨适合我国现况的验证方法，为促进我国医疗大数据的应用和基于医疗保险数据库开展相关研究提供参考。

【关键词】 医疗保险数据库；数据库准确性验证；机器学习；自然语言处理；数据库链接
基金项目：国家自然科学基金（91646107）；国家自然科学基金青年科学基金（81502884）
DOI：10.3760/cma.j.issn.0254-6450.2019.10.027

An overview of validation methods based on the medical claims database

Feng Jingnan, Wang Shengfeng, Zhan Siyan
Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China
Corresponding author: Zhan Siyan, Email: siyan-zhan@bjmu.edu.cn

【**Abstract**】 Medical claims database is an important source of data for studying the characteristics, and burden of diseases, to provide a basis for the development of policy on management. The database is usually used to identify patients through International Classification of Diseases and free text-building algorithms, thus it is crucial to validate whether the algorithm is correctly identifing the targeted population. This paper introduces both traditional and emerging validation methods including machine learning, natural language processing and database linkage *etc*.. We also have tried to present a suitable validation method for the current situation in China, so as to promote the application of big data in medical areas and to provide reference for epidemiology studies, based on medical claims database in this country.

【**Key words**】Medical claims database; Validation; Machine learning; Natural language processing; Database linkage
Fund programs: National Natural Science Foundation of China (91646107); National Natural Science Foundation of China Youth Science Foundation (81502884)
DOI:10.3760/cma.j.issn.0254-6450.2019.10.027

医疗保险数据库（medical claims database）通常由政府或医疗服务机构收集，可以较低收集成本覆盖不同地区的大量人口[1]，同时收集包括患者的人口统计学信息（姓名、出生日期、性别等）、就诊日期、诊断信息以及服务费用等多项信息[2]。与其他类型的数据相比，医疗保险数据的优势在于回忆偏倚小，可精确估计发病率、患病率，以及研究结论对一般人群具有代表性[3]。目前，医疗保险数据库越来越多地被用于评价医疗保健服务的功能、药物的安全性和有效性以及真实世界的观察性研究[4-5]。

但是，由于医疗保险数据库收集数据是为管理而非临床研究目的，且不同级别的医疗机构和地区之间对疾病的诊断水平和诊断编码的要求不同，导致不同地区间的数据质量存在差异[6]，必须对数据收集的准确性进行评估，以便更加可靠地用于临床研究。因此，医疗保险数据库需要对感兴趣的结局事件进行验证，根据选定的参考标准判断数据库内容是否正确[7]，有助于评估从数据库中获得的变量和研究结果的潜在偏倚[8]。

近年来，美国和欧洲地区由于医疗保险数据库发展较早，相关数据准确性验证方法也已较为成熟。我国目前正处于医疗大数据蓬勃发展的时期，相关领域的研究较少。本文旨在通过系统梳理国外医疗保险数据库准确性的验证方法，借鉴欧美等国的经验，并结合我国实际情况，为未来国内的医疗保险数据库准确性验证工作提供参考。

一、传统的数据库准确性验证方法

传统的验证研究方法是直接审阅数据库中的临床病历，这一方法对于评估算法识别目标疾病的能力，验证特定的个

体是否罹患某种疾病或暴露于某种危险因素非常重要[9]。美国食品药品监督管理局(Food and Drug Administration, FDA)“哨点计划”工作组通过系统回顾260篇涉及医疗保险数据库准确性验证的研究发现,最常见的验证方法是将医疗病历作为参考标准,并将阳性预测值(positive predictive values, PPV)作为判断诊断准确性的指标[10]。结合FDA“哨点计划”和美国Medicare/Medicaid医疗保险数据库的病历审查(chart review)[9,11],审查病历信息的方法可分为5个步骤:第一步是验证研究的启动,包括制定工作计划、检索病历的程序和流程,与数据合作伙伴和病历提供方达成合作等。第二步是确定进行病历审查的目标疾病的编码、算法以及样本量,算法通常根据国际疾病分类编码(International Classification of Diseases, ICD)构建。第三步是病历检索。第四步是病历信息提取和判定,通常由2名经过培训的专业人员分别提取每1份病历的信息,提取的信息包括人口统计学信息、诊断信息、诊断编码等。对提取信息的判定,可根据研究结果的复杂程度,由1~2名相关方向的临床专家进行裁定[12]。最后一步是计算PPV,将经过病历审查最终判断为目标疾病的病历数量与进行病历审查的样本量相比,得到PPV,“哨点计划”认为PPV>70%,算法具有较高的可信度[13]。

二、新兴的数据库准确性验证方法

尽管直接审查病历信息的方法行之有效,但这种验证方法耗时耗力。“哨点计划”的1项研究发现,平均每个病历审查研究需要401份病历,每份病历平均80页,仅前3步就占病历审查总费用的65%~81%[9];其次,获得病历需要得到医院、诊所的同意,患者隐私和数据安全需要研究人员和医疗管理机构实施额外的保护措施,不仅增加临床人员的工作量,还会增加研究成本。因此,探索医疗保险数据库准确性验证研究的新方法十分必要[14]。

1. 传统病历审查验证方法的改进:“哨点计划”工作组在1项研究中,通过回顾5项已完成的病历审查验证研究,根据实施步骤将影响病历审查研究效率的因素总结为3条,并提出了针对性改进措施[9]。

第一,在研究启动阶段与数据合作伙伴或病历提供方签订合同时,影响成本和效率的因素主要在于合同内容的不确定性而增加预算和为遵守对方隐私、法律和其他监管要求所耗费的时间和精力。对应改善措施包括:①降低成本:选择成本低的病历提供方;开发病历审查资源需求评分工具(chart review resource intensity score),在研究开展的前期,评估目标疾病开展病历审查所需的病历数和费用;建立与数据合作方的标准预算条目;②隐私保护:提高双方工作范围和操作程序的标准化,建立与数据合作伙伴的数据共享指南,在达到数据隐私的保护要求下,最大限度地减少对病历数量的需求。

第二,在研究开展阶段构建疾病算法时,由于不同的疾病,需要不同的编码检索病历,不同来源的数据库中程序和代码的不一致是主要的限速步骤。对应改善措施包括开发标准化、模块化程序,规范病历审查请求的格式;统一通用数据模型中的疾病编码和字段,使用机器学习等新兴技术构建算法。

第三,在病历检索、信息提取和判定时,主要的问题在于时间过长,特别是当目标疾病需要提取的病历数量或病历中的变量较多且复杂时,不仅耗费大量人力,病历提供方也需耗费大量时间满足研究需求。对应改善措施包括规范标准化操作程序,优化目标疾病需要提取的病历和变量的数量;建立信息提取和判定的人员设置标准,对于定义简单或单独提取、判定不易出错的目标疾病可采用单人进行;建立与数据合作伙伴和病历提供方的通信准则,以确保在病历审查过程中及时反映问题;以及与数据合作伙伴或病历提供方达成协议,尽可能利用已完成的验证研究中的病历信息,以减少病历审查的成本和时间。

2. 机器学习(machine learning)和自然语言处理(natural language processing, NLP)技术应用于数据库准确性验证:机器学习指应用于分类和预测的算法,包括多变量logistic回归、支持向量机、随机森林、NLP等[15]。随着医疗病历的电子化,机器学习应用于数据库准确性验证体现在两方面:一是通过分类技术在病历信息判定的过程中替代人工专家判定;二是基于NLP技术,在算法中增加对非结构化自由文本的处理,并应用于病历信息提取的过程中。

(1)基于机器学习的疾病分类:近年来,已有多项研究应用机器学习模型准确地对目标疾病和暴露因素进行了分类,例如,阿尔茨海默病、中风、耐药性癫痫等[16-18]。以Bergquist等[19]开展的应用机器学习,对医疗保险数据中肺癌严重程度分类的研究为例进行说明。

该研究应用美国癌症注册和Medicare链接的数据库,纳入6个月内接受过≥1次注射或口服化疗药物的肺癌患者,以癌症注册数据中的肺癌分期为“金标准”,通过建立决策树对肺癌严重程度进行分类。研究结局为“早期阶段”(肿瘤Ⅰ~Ⅲ期)和“晚期阶段”(肿瘤Ⅳ期)。协变量分为3类:第1类由临床指南中定义的7个临床指标构成(如肺癌化疗、肺切除手术);第2类包含来自医疗保险数据库的95个人口统计学、治疗和合并症变量(如化疗药物、住院或门诊治疗);第3类则由13个涉及肺癌类型和二级恶性肿瘤的诊断编码构成。分类规则包括“基于最可能”的分类以及“基于预测概率的百分位数分布”的分类。通过以上变量构建不同的算法,计算灵敏度、阳性预测值、ROC曲线下面积。研究结果表明,与仅使用临床变量相比,使用与医疗保险数据扩展的变量和算法,结合机器学习技术,可用于对接受化疗药物的肺癌患者严重程度进行分类,且灵敏度(93%)和准确度(93%)均较高[19]。

(2)基于NLP的非结构化数据处理:病历中包含的很多信息,例如实验室、病理、放射诊断报告,入院、出院摘要以及患者主诉等均是非结构化或半结构化的文本,使得在病历审查人工提取信息时,消耗大量时间和资源。NLP技术是1项应用于人类语言智能处理的方法,不仅可以在提取病历信息时辅助人工,还可在确定目标疾病的算法时,帮助在算法中

加入自由文本的诊断内容，更加全面地发现数据库中的患者并减小偏倚。英国的1项研究表明，在仅有编码的算法中添加文本，可显著提高识别病历的速率和准确性[20]。目前，在电子医疗病历（electronic medical records，EMR）中，已有应用NLP技术针对慢性非传染性疾病、传染病、心理障碍和伤害开发的基于疾病编码和自由文本识别目标疾病的算法[21]，美国疫苗不良事件报告系统（Vaccine Adverse Event Reporting System）也开发了类似的针对疫苗不良事件的算法[22]。以基于“迷你哨点计划”数据，应用NLP对过敏反应非结构化数据进行分类的研究为例进行说明[23-24]。

该项研究的样本是基于既往研究提取的62个患者的病历，根据Sampson对过敏反应的定义，仅检索ICD-9-CM编码995.0和999.4的病历。这些病历经过判定后分为过敏反应（33人）、非过敏反应（27人）和不确定（2人），以此为“金标准”，使用2种NLP方法对上述病历进行分类。2种方法均应用Brighton协作组织（Brighton Collaboration，BC）对过敏反应的定义，与Samposon定义类似，主要区别在于，BC允许不同水平的证据，并且不对患者的已知过敏原做出假设。第1种方法在BC定义的基础上，以与使用肾上腺素有关的关键特征字段作为主要标准构建算法，称为“基于规则”的算法。第2种方法称为“基于相似性”的算法，以BC过敏反应定义中的同义词为标准，通过与自由文本进行比较确定分类：当相似性得分＞0时，归类为过敏反应；相似性得分＝0时，判定为非过敏反应。对于2种方法，分别计算灵敏度、阳性预测值和*F*值（灵敏度和阳性预测值的调和均数）。结果显示，“基于规则”和“基于相似性”算法的性能相似（灵敏度：100% *vs.* 100%，阳性预测值：60.3% *vs.* 57.4%，*F*值：0.753 *vs.* 0.729），NLP可应用于“迷你哨点计划”中对过敏反应文本的分类[23-24]。

3. 链接电子数据库应用于数据库准确性验证：为了解决传统病历审查昂贵、耗时的问题，另一个方案是寻找可以替代病历的“金标准”。一些研究使用出院摘要、患者自发报告、调查问卷等作为“金标准”进行验证，但这些数据并不全面和真实[2]，最好的方式是应该将医疗保险数据库和包含目标疾病真实病历的电子数据库链接，例如注册数据库（registry）、EMR、区域医疗数据库等。

图1中，字母表示各个数据库中包含的患者信息，使用

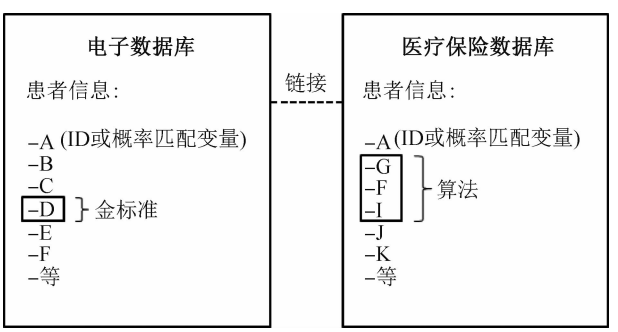

图1　电子数据库链接医疗保险数据库准确性验证流程示意图

G、F、I变量构建算法，在医疗保险数据库中识别目标疾病患者，然后将这些患者通过个人身份识别信息或1组概率匹配变量（probabilistic matching variables）A链接到电子数据库，电子数据库中包含的目标疾病信息D作为“金标准”，以此评估算法识别目标疾病患者的能力。

目前，已有多篇文献应用此法进行数据库准确性验证，例如“哨点计划”工作组总结出6个优先进行验证的疾病：自杀、Ⅰ型糖尿病、高血压危象、肺纤维化、肺动脉高压和自然流产，并为这6个疾病找到最佳的匹配注册数据库，分别为：国家死亡指数数据库（National Death Index）、Ⅰ型糖尿病注册数据库（T1D Exchange Registry）、健康维护组织研究网络高血压注册数据库（Health Maintenance Organization Research Network Hypertension Registry）、宾夕法尼亚特发性肺纤维化注册数据库（Pennsylvania Idiopathic Pulmonary Fibrosis Registry）、评估早期和长期肺动脉高血压疾病管理注册数据库（Registry to Evaluate Early And Long-Term Pulmonary Artery Hypertension Disease Management）、美国CDC卫生统计中心的胎儿死亡数据和全国儿童研究数据库（the Fetal Death Dataset from the CDC National Center for Health Statistics and National Children's Study）[13]。

不同的研究根据数据库变量信息的特点，选择不同的链接方式。在一些电子数据库，当患者的ID或包含个人识别信息的变量可以使用时，数据库之间可通过匹配ID或涉及个人信息的变量直接链接，例如：加拿大安大略省一项在医疗保险数据库中，验证糖尿病算法准确性的研究，即使用患者ID将安大略省的医疗保险数据链接到初级卫生保健EMR数据库中[25]；法国一项验证淋巴瘤算法准确性的研究，通过姓氏、出生姓名、第一姓名、出生日期、性别、出生地6个变量将法国国家医疗保险数据库（French Health Insurance System Database）和塔恩省的癌症注册数据库（Tarn Cancer Registry）链接[26]。但是，当缺少患者身份识别信息或由于保护患者隐私不能使用时，需要使用概率匹配的方法进行链接，即联合使用数据库中的多个变量生成一个新的变量，这些变量在数据库中单独使用时不能识别出唯一的患者，但联合使用时能在最大程度上识别到唯一个体，例如：中国台湾地区在1项验证用于识别急性缺血性卒中、短暂性脑缺血发作或脑内出血患者卒中危险因素的算法准确性研究中，通过4个概率匹配变量（性别、出生日期、入院日期和出院日期）将注册数据库与医疗保险数据库链接，排除注册数据库匹配了多个医疗保险数据库患者的病历[2]；同样，美国1项验证前列腺癌算法准确性的研究，将密歇根州医疗保险数据库中的前列腺癌患者，通过出生日期、前列腺活检日期和泌尿科医生3个概率匹配变量与密歇根泌尿外科改善协作注册数据库（Michigan Urologic Surgery Improvement Collaborative Registry）进行了链接[27]。

三、我国数据库准确性研究现状

2015年8月国务院发布了《促进大数据发展行动纲要》[28]，随后国家发展和改革委员会下发《关于组织实施促进

大数据发展重大工程的通知》[29]，均强调要大力推动政府部门数据共享，推动公共数据资源开放。截至2016年底，我国基本医疗保险参保人数超过13亿人，参保覆盖率稳固在95%以上[30]，是研究我国人群疾病特征、疾病负担、制定管理政策的宝贵资源。此外，EMR数据库、注册数据库和区域医疗数据库发展迅速，为开展数据库准确性验证提供条件[31]。

但是，目前利用医疗保险数据库等大数据资源的验证研究多集中在欧美等国，特别是“哨点计划”在2014年已通过系统回顾有关数据库准确性验证的文献，完成了对16个目标疾病算法的评估[32]。国内对数据库准确性的验证研究较少，Xu等[33]使用2010—2014年首都医科大学附属佑安医院的EMR数据，以病历审查为“金标准”，验证了40种疾病的算法。对于医疗保险数据库准确性的验证研究，一项对亚太地区10个国家或地区的医疗保险数据库准确性验证的系统综述研究显示，在纳入的43篇研究中，我国大陆地区未见到有相关研究[34]。

由于传统的病历审查耗时耗力，并且不利于保护患者的隐私，借鉴欧美等国的经验并结合我国医疗保险数据库的特点，提出适合我国数据库准确性验证的方法，即通过与EMR数据库、注册数据库和区域医疗数据库等将患者的信息进行链接，通过计算PPV验证算法。在实际实施时需要注意以下几点：①患者的信息安全。医疗保险数据库或注册数据库中包含患者的信息，应当受到保护，使用这些数据时应当遵守我国的相关法律和伦理规范；其次，链接所用的涉及到个人身份信息的变量应是去识别化的，链接的变量应在链接后删除[35-36]；②疾病编码与文本构建算法。与欧美等国不同的是，由于在我国临床实际中，医生的处方仍以文字为主，故对于目标疾病的算法，应同时包括疾病编码和临床上所有可能的目标疾病的疾病名称，运用NLP技术以尽可能在医疗保险数据库中抓取到所有的患者；③数据标准化与统一。患者的疾病编码通常由医院信息科工作人员根据医生的诊断进行编码，由于一种疾病可能有多种写法，当临床医生书写不规范时，信息科工作人员可能会因为缺乏相应的临床知识，导致在编码时出现错误。建议在临床实际中，首先要增加医生检索、获得疾病编码的便捷性，由医生在医院信息系统中直接对患者的患病情况进行编码。此外，与欧美等国在构建算法时，使用ICD-10四位编码不同，我国各地疾病编码的不统一也给医疗保险数据的进一步应用带来了困难。我国自2017年2月开始实施《中华人民共和国国家标准疾病分类与代码》，这套编码是根据ICD-10，结合我国国情编制而成的6位编码[37]，但部分地区或省份使用的仍为地方版本的编码，因此，给研究全国性的疾病特征和分布构建算法时带来困难，建议在全国范围内推行统一的疾病编码。

四、展望

医疗保险数据库蕴藏着丰富的信息，是研究我国人群疾病特征、疾病负担、制定管理政策的重要来源。数据库准确性的验证对判断算法是否能正确识别目标疾病人群十分重要，通过疾病编码和诊断名称结合构建目标疾病的算法，然后将患者链接到EMR数据库、注册数据库或区域医疗数据库等进行数据库准确性验证的思路，现阶段可以作为一种尝试方向，为推动我国医疗大数据的应用奠定基础。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] Rimland JM, Abraha I, Luchetta ML, et al. Validation of chronic obstructive pulmonary disease (COPD) diagnoses in healthcare databases: a systematic review protocol[J]. BMJ Open, 2016, 6(6): e011777. DOI: 10.1136/bmjopen-2016-011777.

[2] Sung SF, Hsieh CY, Lin HJ, et al. Validation of algorithms to identify stroke risk factors in patients with acute ischemic stroke, transient ischemic attack, or intracerebral hemorrhage in an administrative claims database[J]. Int J Cardiol, 2016, 215: 277-282. DOI: 10.1016/j.ijcard.2016.04.069.

[3] Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data[J]. J Clin Epidemiol, 2011, 64(8): 821-829. DOI: 10.1016/j.jclinepi.2010.10.006.

[4] Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics[J]. J Clin Epidemiol, 2005, 58(4): 323-337. DOI: 10.1016/j.jclinepi.2004.10.012.

[5] Kern DM, Davis J, Williams SA, et al. Validation of an administrative claims-based diagnostic code for pneumonia in a US-based commercially insured COPD population[J]. Int J Chron Obstruct Pulmon Dis, 2015, 10: 1417-1425. DOI: 10.2147/COPD.S83135.

[6] Cunningham CT, Cai P, Topps D, et al. Mining rich health data from Canadian physician claims: features and face validity[J]. BMC Res Notes, 2014, 7: 682. DOI: 10.1186/1756-0500-7-682.

[7] Lacasse Y, Daigle JM, Martin S, et al. Validity of chronic obstructive pulmonary disease diagnoses in a large administrative database[J]. Can Respir J, 2012, 19: e5-9. DOI: 10.1155/2012/260374.

[8] Thigpen JL, Dillon C, Forster KB, et al. Validity of international classification of disease codes to identify ischemic stroke and intracranial hemorrhage among Individuals with Associated Diagnosis of Atrial Fibrillation[J]. Circ Cardiovasc Qual Outcomes, 2015, 8(1): 8-14. DOI: 10.1161/CIRCOUTCOMES.113.000371.

[9] Sentinel. Sentinel medical chart review gap analysis report[EB/OL]. (2018-04-18)[2018-12-10]. https://www.sentinelinitiative.org/sentinel/methods/sentinel-medical-chart-review-gap-analysis.

[10] McPheeters ML, Sathe NA, Jerome RN, et al. Methods for systematic reviews of administrative database studies capturing health outcomes of interest[J]. Vaccine, 2013, 31 Suppl 10: K2-6. DOI: 10.1016/j.vaccine.2013.06.048.

[11] Centers for Medicare & Medicaid Services. EQR protocol 4 validation of encounter data reported by the MCO[EB/OL]. (2012-09-01)[2018-12-10]. https://www.medicaid.gov/medicaid/quality-of-care/downloads/eqr-protocol-4.pdf.

[12] Cutrona SL, Toh S, Iyer A, et al. Validation of acute myocardial infarction in the food and drug administration's mini-sentinel program[J]. Pharmacoepidemiol Drug Saf, 2013, 22: 40-54. DOI: 10.1002/pds.3310.

[13] Mini-Sentinel. Alternative methods for health outcomes of interest validation [EB/OL]. (2013-08-31) [2018-12-10]. https://www.sentinelinitiative.org/sites/default/files/surveillance-tools/validations-literature/Mini-Sentinel-Alternative-Methods-for-Health-Outcomes-of-Interest-Validation_0.pdf.

[14] Williamson T, Miyagishima RC, Derochie JD, et al. Manual review of electronic medical records as a reference standard for case definition development: a validation study[J]. CMAJ Open, 2017, 5: E830-833. DOI: 10.9778/cmajo.20170077.

[15] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning[J]. Stat Med, 2010, 29: 337-346. DOI: 10.1002/sim.3782.

[16] Magnin B, Mesrob L, Kinkingnéhun S, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI[J]. Neuroradiology, 2009, 51(2): 73-83. DOI: 10.1007/s00234-008-0463-x.

[17] Asadi H, Dowling R, Yan B, et al. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy[J]. PLoS One, 2014, 9(2): e88225. DOI: 10.1371/journal.pone.0088225.

[18] An S, Malhotra K, Dilley C, et al. Predicting drug-resistant epilepsy — a machine learning approach based on administrative claims data[J]. Epilepsy Behav, 2018, 89: 118-125. DOI: 10.1016/j.yebeh.2018.10.013.

[19] Bergquist SL, Brooks GA, Keating NL, et al. Classifying lung cancer severity with ensemble machine learning in health care claims data[J]. Proc Mach Learn Res, 2017, 68: 25-38.

[20] Tate AR, Martin AGR, Murray-Thomas T, et al. Determining the date of diagnosis — is it a simple matter? The impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK primary care[J]. BMC Med Res Methodol, 2009, 9: 42. DOI: 10.1186/1471-2288-9-42.

[21] Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review[J]. J Am Med Inform Assoc, 2016, 23(5): 1007-1015. DOI: 10.1093/jamia/ocv180.

[22] Botsis T, Buttolph T, Nguyen MD, et al. Vaccine adverse event text mining system for extracting features from vaccine safety reports[J]. J Am Med Inform Assoc, 2012, 19(6): 1011-1018. DOI: 10.1136/amiajnl-2012-000881.

[23] Ball R, Toh S, Nolan J, et al. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA Sentinel System[J]. Pharmacoepidemiol Drug Saf, 2018, 27(10): 1077-1084. DOI: 10.1002/pds.4645.

[24] Walsh KE, Cutrona SL, Foy S, et al. Validation of anaphylaxis in the Food and Drug Administration's Mini-Sentinel [J]. Pharmacoepidemiol Drug Saf, 2013, 22(11): 1205-1213. DOI: 10.1002/pds.3505.

[25] Lipscombe LL, Hwee J, Webster L, et al. Identifying diabetes cases from administrative data: a population-based validation study[J]. BMC Health Serv Res, 2018, 18: 316. DOI: 10.1186/s12913-018-3148-0.

[26] Conte C, Palmaro A, Grosclaude P, et al. A novel approach for medical research on lymphomas: a study validation of claims-based algorithms to identify incident cases[J]. Medicine, 2018, 97(2): e9418. DOI: 10.1097/MD.0000000000009418.

[27] Modi PK, Kaufman SR, Qi J, et al. National trends in active surveillance for prostate cancer: validation of medicare claims-based algorithms[J]. Urology, 2018, 120: 96-102. DOI: 10.1016/j.urology.2018.06.037.

[28] 国务院. 促进大数据发展行动纲要[J]. 成组技术与生产现代化, 2015, 32(3): 51-58. DOI: 10.3969/j.issn.1006-3269.2015.03.012.
State Council. Platform for actio of big-data development [J]. Group Tech Product Modernization, 2015, 32(3): 51-58. DOI: 10.3969/j.issn.1006-3269.2015.03.012.

[29] 国家发展和改革委员会办公厅. 关于组织实施促进大数据发展重大工程的通知[EB/OL]. (2016-01-07) [2018-12-20]. http://bigdata.sic.gov.cn/archiver/bigdata/UpFile/Files/Default/20160603145033968072.pdf.
General Office of the National Development and Reform Commission. Notification on organizing and implementing the major project for big-data development [EB/OL]. (2016-01-07) [2018-12-20]. http://bigdata.sic.gov.cn/archiver/bigdata/UpFile/Files/Default/20160603145033968072.pdf.

[30] 国务院新闻办公室.《中国健康事业的发展与人权进步》白皮书[EB/OL]. (2017-09-29) [2018-12-20]. http://www.scio.gov.cn/ztk/dtzt/36048/37159/37161/Document/1565175/1565175.htm.
State Council Information Office of China. White paper for health development and human rights progress in China [EB/OL]. (2017-09-29) [2018-12-20]. http://www.scio.gov.cn/ztk/dtzt/36048/37159/37161/Document/1565175/1565175.htm.

[31] 杨羽, 詹思延. 上市后大数据药品安全主动监测模式研究的必要性和可行性[J]. 药物流行病学杂志, 2016, 25(7): 401-404.
Yang Y, Zhan SY. Analysis of necessity and feasibility in studies of post-marketing drug safety active surveillance based on big data[J]. Chin J Pharmacoepidemiol, 2016, 25(7): 401-404.

[32] Mini-Sentinel. 16 health outcomes of interest for surveillance preparedness [EB/OL]. (2014-07-01) [2018-12-10]. https://www.sentinelinitiative.org/sentinel/surveillance-tools/validations-lit-review/16-health-outcomes-interest-surveillance.

[33] Xu Y, Li N, Lu MS, et al. Development and validation of method for defining conditions using Chinese electronic medical record [J]. BMC Med Inform Decis Mak, 2016, 16: 110. DOI: 10.1186/s12911-016-0348-6.

[34] Koram N, Delgado M, Stark JH, et al. Validation studies of claims data in the Asia-Pacific region: a comprehensive review [J]. Pharmacoepidemiol Drug Saf, 2019, 28(2): 156-170. DOI: 10.1002/pds.4616.

[35] Palmieri L, Veronesi G, Corrao G, et al. Cardiovascular diseases monitoring: lessons from population-based registries to address future opportunities and challenges in Europe[J]. Arch Public Health, 2018, 76: 31. DOI: 10.1186/s13690-018-0283-3.

[36] Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research[J]. J Epidemiol Community Health, 2014, 68: 283-287. DOI: 10.1136/jech-2013-202744.

[37] 国家卫生和计划生育委员会. GB/T 14396—2016 疾病分类与代码[S]. 北京: 中国标准出版社, 2017.
National Health and Family Planning Commission. GB/T 14396-2016 Classification and codes for disease [S]. Beijing: Standards Press of China, 2017.

（收稿日期：2019-02-22）
（本文编辑：李银鸽）