· 方法学指南解读 ·

药物流行病学研究中的二次数据源

卓琳[1#]　马慧宁[2#]　贾敏[3]　杨羽[4]　孙凤[5]　宫建[3]　詹思延[1,5]

（1. 北京大学第三医院临床流行病学研究中心　北京 100191；2. 宁夏医科大学总医院心脑血管病医院；
3. 沈阳药科大学药物流行病与临床药物评价课题组；4. 北京大学健康医疗大数据国家研究院；
5. 北京大学公共卫生学院流行病与卫生统计学系）

摘　要　真实世界数据在药品与医疗器械临床评价中应用越来越受到重视，通过反映实际诊疗状况从而为医疗决策提供依据，已成为国内外学者的研究热点。电子健康数据的发展和普及为开展药物流行学研究提供了多样的数据基础，但由于收集目的不同造成各类数据各自不同的特点，在进行研究时需了解数据库特性谨慎选择适合的数据源。本文以《中国药物流行病学研究方法学指南》为基础，从数据源和数据收集方法概述、国内外数据源现状、存在的问题、相应解决方案、案例分析等方面进行系统综述，介绍药物流行病学研究中二次数据及数据收集方法的特点，举例介绍各国进展，并简要分析其优缺点和研究中所需注意的事项以及对我国药物流行病学研究的启示。

关键词　药物流行病学；数据源；二次数据；电子病历；医疗保险数据

中图分类号：R181.3^{+}5　　**文献标识码**：A　　**文章编号**：1005-0698(2021)04-0219-05

Secondary Data Sources in Pharmacoepidemiology Studies

Zhuo Lin[1], Ma Huining[2], Jia Min[3], Yang Yu[4], Sun Feng[5], Gong Jian[3], Zhan Siyan[1,5]

1. Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing 100191, China; 2. Cardiovascular and Cerebrovascular Disease Hospital, Ningxia Medical University General Hospital; 3. Research Group of Jian Gong on Pharmacoepidemiology and Clinical Drug Evaluation, Shenyang Pharmaceutical University; 4. National Institute of Health Data Science, Peking University; 5. Department of Epidemiology and Bio-statistics, School of Public Health, Peking University

ABSTRACT　The application of real-world data in the clinical evaluation of medicines and medical devices has received more and more attention. It has become a research hotspot for scholars at home and abroad by reflecting the actual diagnosis and treatment status for medical decision-making. Electronic health data development and popularization offer a diversified data basis for conducting pharmacoepidemiological research, but it is still necessary to carefully select the appropriate data sources. Based on the *Guide on Methodological Standards in Pharmacoepidemiology*, this overview systematically summarized the secondary data source and data collection method for pharmacoepidemiology researches. The advantages and disadvantages of these data are analyzed briefly and the matters needing attention in the study.

KEY WORDS　Pharmacoepidemiology; Data source; Secondary data; Electronic health records; Claims data

药物流行病学是指采用流行病学的研究方法研究药物在人群中的使用情况和使用效果的学科[1]。药物流行病学不仅能补充完善药品上市前临床研究所获得的信息，还可以获取药品上市后临床研究未曾获得的新信息。随着药物流行病学研究方法的发展，研究者对如何合理选择数据以回答研究问题更为关注。同时，随着医疗信息化的飞速发展和医疗数据的快速积累，利用海量真实世界数据开展药物流行病学研究成为可能。药物流行病学数据源根据来源不同可分为一次数据和二次数据，其中二次数据源是指为其他目的（如临床诊疗、药品管理等）而收集的数据，常用的为电子健康数据库[2]。相比纸质医疗记录，在数据存储和查询方面更有优势的电子健康档案为药物流行病学研究提供了详细的患者信息。不仅如此，大数据信息技术的进步，也使得链接多个结构特异的数据源，甚至利用网络及社交媒体中患者数据成为可能。本文简要介绍药物流行病学研究中常用的二次数据源及部分重点数据库，并

#同等贡献

基金项目：北京大学医学部教育教学研究课题（编号：2020YB03）

通信作者：宫建　Tel：15840064816　E-mail：fanxing1230@163.com

詹思延　Tel：(010)82805162　E-mail：siyan-zhan@bjmu.edu.cn

举例说明不同数据的应用和优缺点。

1　医疗保险数据

医疗保险数据库是为管理保险、提供药房和医疗报销记录目的而建立的数据库，记录了参保用户的基本信息、就诊医师或科室信息、医疗记录（门诊和住院）、用药及其费用等信息[3]。记录详实的数据库甚至有可能记录了患者加入医疗保险及退出的日期。这种数据库具有涵盖人群广，收集数据时间长的特点，比随机对照试验（random control trial，RCT）更能代表真实世界的医疗实践。

医疗保险数据库可以用于不同研究目的：追踪医疗资源使用情况；通过体现临床日常诊疗行为，为真实世界有效性研究提供相关证据；用于观察罕见事件；可以通过筛选新药不良事件及其发生的危险因素进行上市后安全性研究[4]。此外，也可以利用这些数据评估卫生政策变化的影响以及卫生资源利用与药物经济学研究。

医疗保险数据库的来源根据医疗保险机构的不同，主要分为商业保险（commercial insurance）、公共保险（public insurance）和区域性保险数据库。

1.1　商业保险

商业保险公司一般均设有投保者专有数据库，如美国蓝盾医疗保险（Blue Cross Blue Shield），Aetna 保险或 United Healthcare 保险等，研究者可以向保险公司申请，并支付一定使用费获取数据使用权。此外还有常用于药物流行病学研究的 IMS LifeLink 医疗保险整合数据库[5]以及 MarketScan 数据库[6]等，涵盖人群均已超过 100 万，其中 MarketScan 数据已包含 2.65 亿患者的脱敏诊疗数据[7]。

1.2　公共保险

公共保险一般由各个国家和地区政府部门设立，由于各个国家医疗体系及报销体系不同，公共保险形式也不尽相同。

（1）美国：如在美国以联邦医疗救助数据库（Medicaid）、联邦医疗保险（Medicare）、美国退伍军人管理局（Verterans Health Administration）数据库为主，其中 Medicaid 和 Medicare 数据库覆盖了 1 亿多美国人口[8]。Medicare 是美国全国范围内老年和残障人口基础医疗保险数据库，包括几乎所有 65 岁以上人群、符合条件的伤残人群以及晚期肾病患者的入院治疗、门诊治疗、专业护理机构、家庭保健及临终关怀机构的费用，由受保人的住院（A 部分）、补充性（B 部分）医疗记录以及处方药物数据库（D 部分）等组成。研究者可向美国研究数据援助中心（Research Data Assistance Center，ResDAC）申请使用 Medicare 5% 抽样数据，但仅限于研究用途。Medicaid 是美国州政府的医疗救助数据库，包含在每个州计划下的向医疗救助受益人提供的如药房、住院、门诊、实验室化验等所有服务信息。Medicare 和 Medicaid 均可以相互关联或者关联到其他数据源。将单独投保 Medicare 或 Medicaid，以及 Medicare 和 Medicaid 双重投保受益人的报销数据整合形成的慢性病数据仓库（chronic condition data warehouse，CCW）是其中应用较多的数据库。CCW 涵盖了 27 种慢性疾病、25 种精神疾病、以及吸烟状况和身体状况信息，极大地扩充了研究变量。

（2）法国、韩国、加拿大等国家：与美国医疗保险结构不同，法国医疗保险虽然可以分为普通保险、农业类保险和非农业、非受雇、自由职业者保险，但每个单独的保险计划均会汇总在法国国家健康保险数据库（SNIIRAM）中，包含所有报销索赔的匿名信息，与国家医院出院摘要数据库系统（national hospital-discharge summaries database system，PMSI）和国家死亡登记处相关联[9]。此外，韩国医疗保险数据库[10]利用上述地区医疗保险数据库开展了多项研究，覆盖了药物流行病学研究的各个方面。

在加拿大，医疗保险数据库以区域性为主，如萨斯喀彻温省健康数据库、曼尼托巴省健康研究数据库等[11]，均是基于省级行政区划的健康保健医保数据库，涵盖人口已超过百万，涉及死亡率、药房、住院、门诊及其他医疗保险数据。

（3）中国：我国目前以基本医疗保险（包括职工基本医疗保险和城乡居民基本医疗保险）为主，补充商业保险为辅，医疗保险数据库建立起步时间较晚，加之汇总全国患者数据任务庞大，自 2008 年起根据既定的抽样方法，在中国医疗保险研究会（China Health Insurance Research Association，CHIRA）管理下，利用系统抽样的方法，每年按照 5% 的抽样比例从抽样城市的城镇基本医疗保险参保人群中抽取患者一年内的就诊记录纳入数据库，将各地抽样数据汇总成全国基本医疗保险抽样数据库。从数据类型方面，2013 年前主要以住院患者信息为主，2013 年后逐步纳入门诊患者数据。目前数据包括患者人口学信息，门/急诊诊断、医疗服务、药品处方等记录以及报销比例等信息[12]。目前该医保数据库尚未开放，研究人员需向 CHIRA 等权威机构提出申请，经严格审核后方可进行学术使用。我国台湾地区已

建立健康保险研究数据库(NHIRD),进行了诸多药物利用和药品不良反应分析的研究。此外这种纵向随访时间长的数据库可以作为估算疾病患病率和治疗模式的良好数据来源,近些年随着机器学习数据挖掘技术的发展,关联规则、频繁模式挖掘等技术也应用在医疗保险数据库中[13~16]。

近些年,在国家政府的鼓励下,已有大量运用上述数据库甚至关联数据库进行药物流行病学以及药品上市后研究的研究报道,为药物的有效性、安全性和经济性评价方面提供了丰富的资源。虽然上述医疗保险数据库是药物流行病学研究重要的数据源之一,但仍存在以下不足之处:①在比较两种或多种医疗干预措施及其相应结局时缺乏必要的医疗信息;②仅包含被保患者的诊疗记录,且未纳入医疗保健系统之外的事件或者诊疗行为,普适性受到限制;③对患者既往疾病状态及患者个人信息记录不完整,有些数据库缺乏实验室检查结果和诊断试验相关结果[17]。此外,不同数据库覆盖人群不同,所包含数据变量有所不同,数据质量依赖于对每一项记录的准确编码,研究者必须清楚保险理赔数据的生成过程和标准,谨慎选择数据源。

2　电子健康档案/电子病历

电子健康档案(electronic health records, EHRs)指电子化的居民健康档案,是关于医疗保健对象健康状况的信息资源库,该信息资源库以计算机可处理的形式存在,并且能够安全的存储和传输,各级授权用户均可访问,可以包含生命体征、既往病史、诊断、病程记录、处方信息、过敏信息、实验室检查数据、免疫接种日期和影像报告等。其中诊疗信息称为电子病历(electronic medical records, EMRs)。EMRs 指医务人员在医疗活动过程中,使用医疗机构信息系统生成的文字、符号、图表、图形、数据、影像等数字化信息,并能实现存储、管理、传输和重现的医疗记录,是病历的一种记录形式,相比 EHRs 更容易获取。

EMRs 一般包括患者的人口统计资料、诊断、治疗过程、医嘱、病程、检查结果等信息,从时间跨度上覆盖患者整个生命周期。两者数据存储分为结构化形式(如人口学信息、实验室检查结果等)和非结构化形式(如病程记录、影像学检查描述等)。相比结构化数据,非结构化数据提取难度较大,文本数据的提取目前仍是 EMRs 应用于药物流行病学研究的巨大阻碍。EMRs 主要以电子病历数据仓库、分布式研究网络及区域医疗数据库等形式存在,其中以电子病历数据仓库最为常见。

(1)电子病历企业级数据仓库(enterprise data warehouse, EDW)和临床数据仓库(clinical data repository, CDR)都是医疗机构内最常用的数据架构。2003 年以来,中国中医科学院电子数据仓库整合了北京市、深圳市、山西省、广东省、福建省、河北省、吉林省共 60 余家三级医院的信息系统数据,截止 2015 年已收集 300 万个体数据[18]。

(2)分布式研究网络(distributed research network, DRN)是为不同卫生保健系统间研究而创建的数据架构。DRN 将不同机构,地理位置不同的数据库联系起来,在不影响合作方原始数据及数据所有权的同时,共享有限的、研究相关的数据以回答研究问题。如由 17 名成员组成的美国健康维护组织研究网络(Health Maintenance Organization Research Network, HMORN)虚拟数据仓库(virtual data warehouse, VDW)能够在多个领域进行大规模、基于人群的药物流行病学研究,还可以进行疾病监测,发现罕见事件,并挖掘医疗产品使用模式[19]。PopMedNet 平台[20]整合了肿瘤注册登记与 HMORN VDW 数据,并且支持 HMORN VDW 数据模型、迷你哨点计划(mini-sentinel initiative)通用数据模型和 EHR 公共卫生支持(EHR Support for Public Health, ESP)数据模型[21]。同样,分布式动态治疗研究网络(distributed ambulatory research in therapeutics network, DARTNet)整合了不同来源的独立 EHRs 或医院数据库(如药房、实验室、收费数据),目前已有 12 个国家和地区参与,由 85 家医疗机构、13 个教学医疗中心和 3 000 名临床医生组成,可以申请付费使用部分数据[22]。

此外,为了更好地进行药物流行病学研究,美国以患者为中心的实效研究所(Patient-Centered Outcomes Research Institute, PCORI)整合了临床数据研究网络(Clinical Data Research Networks, CDRNs)和患者强化研究网络(Patient-Powered Research Networks, PPRNs)的数据,创建了一个功能强大、代表性强的研究网络,将其称之为国家以患者为中心的临床研究网络(National Patient-Centered Clinical Research Network, PCORnet)。其中,CDRNs 是两个或多个医疗保健系统的集成网络,PPRNs 则由患者和宣传组织发起,收集患者医疗数据并召集特殊疾病研究人员[23]。

(3)区域医疗数据库:近年来我国在区域卫生

信息平台中采集存储了大量的居民医疗健康数据，如宁波市、天津市、厦门市、济南市、四川省及内蒙古自治区等均已尝试从政府层面整合政务数据和电子健康档案数据。其中建设最早，且最为完善的是宁波市鄞州区的区域医疗数据，已经从基层医疗机构以及二/三级医院信息系统中采集了近 10 年的医疗卫生数据[18]。截止 2015 年底，鄞州区户籍人口电子档案累计建档 122.2 万人，电子健康档案建档率已达 98.19%，规范化居民电子健康档案建档率达到 80%。该平台提供患者基本信息、门诊信息、住院信息、检查检验信息、体检信息。其中处方数据来源于区内所有医疗机构的日常门诊处方和住院医嘱，包含患者信息、科室医生信息、诊断信息、处方信息 4 大类 87 个变量。截止 2015 年，门诊处方数据量已达到 6 358 万条，医嘱处方数量已达到 47 万条，数据规模达到 3.3TB。此外，由于鄞州区人口的流动性较低，且数据收集覆盖全区所有医疗机构，因此能够长时间较完整地记录区内居民的医疗信息，这为研究提供了关键的结局信息，如不良反应、疾病转归、再住院率、医疗花费等。

EHRs/EMRs 在患者信息记录方面更为详细、复杂，往往涉及变量和内容丰富，在利用其应用于研究时需要考虑一下几点：①由于患者接受的医疗保健的不同，其诊疗记录间隔也有所不同；②EHR 系统不同，包含变量也不尽相同，如有些不包括住院患者用药情况；③不同医院或不同地区就诊可能导致就诊信息的中断；④部分 EHRs 未记录急诊治疗内容；⑤由于各个国家和地区数据开放程度不同，需在遵守相关法规/条例的前提下，申请/付费使用数据。

3　共同支付者数据库/关联数据库

顾名思义，共同支付者数据库（multipayer/co-payment databases）指包含了商业和公共保险数据的数据库，通常覆盖百万人群，提供多个保险公司的个体水平医疗保险数据，并且可以付费使用。共同支付者数据库常见于商业保险覆盖程度高的国家，目前逐渐发展不断纳入更多数据来源，形成了包括 EHRs、医疗保险数据库以及注册登记的关联数据库（linked database）。如 Optum Labs 关联了 2 亿患者的医疗保险、医疗记录以及患者报告的健康事件的脱敏数据，提供了患者人口统计、健康行为以及医疗花费等方面的细节信息[24, 25]。此外，临床实践研究数据库（Clinical Practice Research Datalink, CPRD）也是一种关联数据库，它将初级综合卫生保健档案、次级卫生保健数据以及其他数据源关联，形成了包含社区用药、医院内用药、实验室检查、医院疾病编码、健康咨询、疾病注册以及癌症等数据的庞大的数据库。目前，CPRD 覆盖了 600 余家医疗机构超过 5 000 万人口（目前注册患者 1 600 万人），为药物流行病学研究提供纵向、大规模人群观察性数据[26]。

4　社交媒体数据

尽管社交媒体网站获得的数据仍需要进一步验证和研究，但仍然有望用于药物流行病学研究。随着患者对自身健康关注度和对医疗卫生决策参与度越来越高，新型电子社交网络模型、健康消费主义的兴起，互联网上也有越来越多的患者信息，可对其进行汇总将其应用于药物流行病学研究。已有研究利用社交媒体数据用于药物流行病学疗效比较研究及安全性研究[27]，并得到了一定的认可。虽然使用通过社交媒体提供的患者数据尚处于起步阶段，随着信息技术的不断发展，仍然可以是进行药物流行病学研究的宝贵资源。

5　小结

药物流行病学研究的实施需要有多源、大型数据库的支持以回答真实世界研究问题。在药物流行病学研究的电子医疗数据基础设施建设时，主要面临的挑战及机遇包括：①项目组需要花费大量精力和资源建立和维护数据合作方之间共享数据的合作伙伴关系；②研究者需要熟悉数据源的多源异构性、数据质量及互操作性，以及一系列临床信息工具、平台和模型的优缺点，以评估这些大型数据库是否满足研究需求；③因为受保护健康信息（protected health information, PHI）保护等法律和伦理框架在不断发展，还需注意这些数据库的伦理保护和地方管理[28]。

当使用上述数据库尤其是二次数据源时，应注意以下几点：①数据采集的完整性，如数据库是否可信地采集了患者所有的健康记录，数据库的覆盖性、信息完整性、时间长度等方面是否有明显缺欠；②通过数据库评价药物暴露时带来的偏倚；③通过数据库定义结局时的有效性；④数据库彼此之间的不一致性。

目前公开发表的基于单个数据库或多源数据进行的药物流行病学研究数目越来越多，但是这些研究更依赖于现有的、程度不等的数据，以及分析方法的分析能力，研究设计与分析方法各有不同。对此，

中国药学会于2019年发布团体标准——《药物流行病学研究方法学指南》[2]可为我国药物流行病学研究提供规范化指南,与此同时,国内外数据库应用的成功案例也可作为构建我国药物流行病学研究数据库的参考,有效利用我国丰富的医疗数据资源。

参 考 文 献

1 International Society for Pharmacoepidemiology. About pharmacoepidemiology[EB/OL]. (2016-07) [2020-11] https://www.pharmacoepi.org/about-ispe/about-pharmacoepidemiology/

2 中国药学会药物流行病学专业委员会. 中国药物流行病学研究方法学指南[J]. 药物流行病学杂志, 2019, 28(10): 695-700

3 吴胤歆, 林承燊, 杨勇鹏. 医疗保险数据库数据质量验证研究[J]. 中国卫生信息管理杂志, 2010, 7(6): 78-80

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

5 Garland A, Gershengorn HB, Marrie RA, et al. A practical, global perspective on using administrative data to conduct intensive care unit research[J]. Ann Am Thorac Soc, 2015, 12(9): 1373-1386

6 Adamson DM, Chang S, Hansen LG. Health research data for the real world: The Marketscan Databases[R]. New York: Thompson Healthcare, 2008

7 Databases IMR. Data, tools and services options designed for life sciences[EB/OL]. (2020-10) [2020-10-30] https://www.ibm.com/sa-en/products/marketscan-research-databases

8 Centers for Medicare & Medicaid Services. Medicare, medicaid, and chip populations for Cy 2020[EB/OL]. (2020-08-20) [2020-11-08] https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CMS-Fast-Facts

9 Bezin J, Duong M, Lassalle R, et al. The National Healthcare System claims databases in France, Sniiram and Egb: Powerful tools for pharmacoepidemiology[J]. Pharmacoepidemiol Drug Saf, 2017, 26(8): 954-962

10 Noh Y, Lee J, Shin S, et al. Antiplatelet therapy of cilostazol or sarpogrelate with aspirin and clopidogrel after percutaneous coronary intervention: A retrospective cohort study using the Korean National Health Insurance Claim Database [J]. PLoS One, 2016, 11(3): e0150475

11 郭剑非, 雷静, 岳晓萌, 等. 如何利用观察性医药数据库进行药物流行病学的安全风险管理研究[J]. 药物流行病学杂志, 2015,24(2): 83-93

12 刁孝华. 我国医疗保障体系的构建时序与制度整合[J]. 财经科学, 2010 (3): 77-84

13 Hsiao FY, Hsieh PH, Huang WF, et al. Risk of bladder cancer in diabetic patients treated with rosiglitazone or pioglitazone: A nested case-control study[J]. Drug Saf, 2013, 36(8): 643-649

14 黄郁庭. 使用时间关系规则侦测台湾人口预期外药物不良反应[D]. 台北:台湾大学硕士学位论文, 2017

15 Kanti BM, Lin CC, Liu CL, et al. Synthesizing electronic health records using improved generative adversarial networks[J]. J Am Med Inform Assoc, 2019,26(3): 228-241

16 Lai CH, Huang LC, Holby SN, et al. Kidney stone history and adverse outcomes after percutaneous coronary intervention[J]. Urology, 2020, 136: 75-81

17 Esposito D, Migliaccio-Walle K, Molsen E. Reliability and validity of data sources for outcomes research & disease and health management programs[M]. International Society for Pharmacoeconomics and Outcomes Research, 2013:223-224

18 Yang Y, Zhou X, Gao S, et al. Evaluation of electronic healthcare databases for post-marketing drug safety surveillance and pharmacoepidemiology in China[J]. Drug Saf, 2018, 41(1): 125-137

19 Ross TR, Ng D, Brown JS, et al. The Hmo Research Network Virtual Data Warehouse: A public data model to support collaboration[J]. Egems, 2014, 2(1): 1049

20 Sittig DF, Hazlehurst BL, Brown J, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data[J]. Med Care, 2012, 50(Suppl): S49

21 Melanie D, Kyle E, Zachary W, et al. Software-enabled distributed network governance: The popmednet experience [J]. Egems, 2016, 4(2): 1213

22 Dartnet Institute. Informing practice improving care[EB/OL]. (2020-11) [2020-10-09] http://www.dartnet.info/default.htm

23 Timbie JW, Rudin RS, Towe V, et al. National patient-centered clinical research network (Pcornet) phase Ⅰ[R]. Santa Monica, CA:RAND,2015

24 Wallace PJ, Shah ND, Dennen T, et al. Optum labs: Building a novel node in the learning health care system[J]. Health Aff(Millwood), 2014, 33(7): 1187-1194

25 Optumlabs. Working to solve health care's greatest challeges [EB/OL]. (2020-11) [2020-10-09] https://www.optumlabs.com/about/story.html

26 Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical practice research datalink (Cprd) [J]. Int J Epidemiol, 2015, 44(3): 827-836

27 Curtis JR, Chen L, Higginbotham P, et al. Social media for arthritis-related comparative effectiveness and safety research and the impact of direct-to-consumer advertising[J]. Arthritis Res Ther, 2017, 19(1): 48

28 Holve E, Segal C, Lopez MH. Opportunities and challenges for comparative effectiveness research with electronic clinical data: A perspective from the Edm Forum[J]. Med Care, 2012,50(7):S11-S18

(2021-02-24 收稿　2021-03-15 修回)