· 药物流行病学理论与方法 ·

观察性健康数据科学和信息学组织实用工具简介

郑轶　郭晓晶　许金芳　迟立杰　郭志坚　陈晨鑫　梁际洲　韦连慧　陈枭　叶小飞　贺佳
（海军军医大学卫勤系军队卫生统计学教研室　上海 200433）

摘　要　运用真实世界大数据进行药品不良反应主动监测受到了国内外学者的高度关注。其中如何实现多个形式不一数据库之间的数据整合，仍然是发挥大数据优势对药品不良反应进行主动监测时存在的一个巨大的挑战。本文旨在通过介绍观察性健康数据科学和信息学（OHDSI）组织的几种工具软件，为我国开展药品不良反应主动监测的数据整合提供思路及方向。

关键词　观察性健康数据科学和信息学；提取-转换-加载；White Rabbit；Rabbit in a hat；Usagi；OHDSI 方法库；ATLAS

中图分类号：R181.3^{+}5　　**文献标识码**：A　　**文章编号**：1005-0698（2021）12-0837-05

DOI：10.19960/j.cnki.issn1005-0698.2021.12.010

An Introduction to OHDSI Tools

Zheng Yi, Guo Xiaojing, Xu Jinfang, Chi Lijie, Guo Zhijian, Chen Chenxin, Wei Lianhui, Chen Xiao, Ye Xiaofei, He Jia
Department of Health Statistics, Naval Military Medical University, Shanghai 200433, China

ABSTRACT　Active monitoring of adverse drug reactions using real-world big data has attracted great attention from scholars at home and abroad. Among them, how to achieve data integration between multiple forms of data, take advantage of big data, and actively monitor adverse drug reactions is still a huge challenge. The purpose of this article is to introduce several OHDSI organization tools to provide ideas and directions for the data integration of active monitor of adverse drug reactions in China.

KEY WORDS　Observational health data sciences and informational; Extraction-transformation-loading; White Rabbit; Rabbit in a hat; Usagi; OHDSI methods base; ATLAS

药品安全关乎人类的健康与生命，也是公共卫生领域的研究重点。药品不良反应（adverse drug reaction，ADR）监测是挖掘 ADR 的重要手段，监测手段主要分为主动监测和被动监测。由于被动监测的数据依赖于自主上报系统（spontaneous reporting system，SRS），存在上报报告信息不全、报告重复、虚假关联等问题，会对所发现的 ADR 信号真实性产生影响[1]。主动监测由于其对信息的收集方式有一定强制性，信息收集得更全面，对风险信号的识别更及时、准确[2]。所以近年来，国内外学者将目光更多地转向了主动监测。

随着电子信息技术不断发展，其在医疗行业得到广泛应用，逐步实现了医疗记录信息化与医疗健康相关的各种数据库建立，伴随着大数据分析技术不断完善，使运用真实世界大数据进行主动监测变成可能[3]。但是如何实现多个形式不一数据库之间的数据整合，仍然是发挥大数据优势对 ADR 进行主动监测时存在的一个巨大的挑战。本文旨在通过介绍观察性健康数据科学和信息学（observational health data sciences and informational，OHDSI）组织的几种工具软件，为我国开展 ADR 主动监测的数据整合提供思路及方向。

1　OHDSI 组织实用工具的开发背景

美国食品药品管理局（Food and Drug Administration，FDA）于 2008 年联合美国药品研究与制造商

基金项目：国家自然科学基金项目（编号：82073671）；上海市自然科学基金项目（编号：18ZR1449500）；军队双重建设项目-03；上海市卫计委优秀青年医学人才培养计划项目（编号：2018YQ47）；上海市公共卫生学科带头人项目（编号：GWV-10.2-XD22）；上海市公共卫生优青计划项目（编号：GWV-10.2-YQ33）

通信作者：贺佳　Tel：（021）81871441　E-mail：hejia63@yeah.net
叶小飞　Tel：（021）81871442　E-mail：yexiaofei@smmu.edu.cn

协会(the Pharmaceutical Research and Manufacturers of America,PhRMA)、美国国家卫生研究院基金会(the Foundation for the National Institutes of Health,FNIH)以及药品生产厂等共同合作成立了观察性医疗结果合作组织(Observational Medical Outcomes Partnership,OMOP)[4]。该组织最初的目的是监测药品的安全性和有效性,逐渐发展为利用观察性健康数据库对药品进行探索。由于该组织的项目重要性得到越来越多学者关注,在其项目结束后,又成立了一个新的项目——OHDSI。OMOP 主要是侧重于方法学方面的研究,而 OHDSI 组织在此基础上,通过开发相应的工具软件,实现对真实世界具体案例进行分析。OHDSI 组织是一个国际合作组织,目前已有来自美国、英国、加拿大等几十个国家、上百个组织参与,旨在通过大规模的分析来发现观察性健康数据的价值[5]。

通用数据模型(common data model,CDM)是 OMOP 的核心[6],其目的是通过建立一套统一的数据标准将不同来源的数据进行标准化,从而实现数据沟通、整合及分析[7,8]。OHDSI 组织在 OMOP 的 CDM 模型以及语言设定之上,构建了一个开放的观察性数据网络平台,通过运用提取-转换-加载(extraction-transformation-loading,ETL)技术,将不同来源的数据转为 CDM 格式的标准化数据,并运用相应的统计分析方法,使得对于涵盖多种数据库的大型研究可以快速、有效地进行。

其中,为了使研究者们更加方便地进行研究,OHDSI 组织针对 ETL 技术,基于 Java 语言开发了 3 款工具软件:White Rabbit、Rabbit in a hat、Usagi,以及针对数据分析开发了两款小工具软件:OHDSI 方法库和 ATLAS。

2　OHDSI 组织的 5 款工具软件介绍

OHDSI 组织的 5 款工具软件的主要功能和优点见表 1。5 款工具软件的工作流程图见图 1。

目前已有许多依据 OHDSI 组织的技术开展的研究,如 2019 年 Suchard 等[9]基于 OHDSI 技术框架,将 490 万例患者数据进行整合,分析了目前处于一线的

表 1　5 款工具软件的主要功能和优点简介

项目	White Rabbit	Rabbit in a hat	Usagi	OHDSI 方法库	ATLAS
主要功能	对源数据进行扫描并创建一份含有必要信息的表格、字段、字段值的 Excel 表格	将 White Rabbit 扫描所生成的必要信息表与 OMOP CDM 中各个表和列进行连接,并生成 Word 文档	辅助手动创建代码映射	是由 OHDSI 组织开发的 R 包,可以对标准化后的数据进行各种分析	是一个基于 Web 界面的交互式分析数据的平台
优点	①支持多种数据格式; ②生成的数据保护患者隐私; ③有一定的数据分析能力	将源数据不同种类的数据与 CDM 匹配,减少 ETL 转换工作量	①如果源数据是非英语,一般会通过谷歌翻译将专业术语转为英语,并创建翻译列; ②有匹配分数,分数低的可由人工进行二次判断	①可以快速调用,避免重投开始编写代码; ②这些 R 包不仅提供了数据分析的功能,而且还提供了一定的制表画图的功能	便于不善于编码的科研人员数据分析

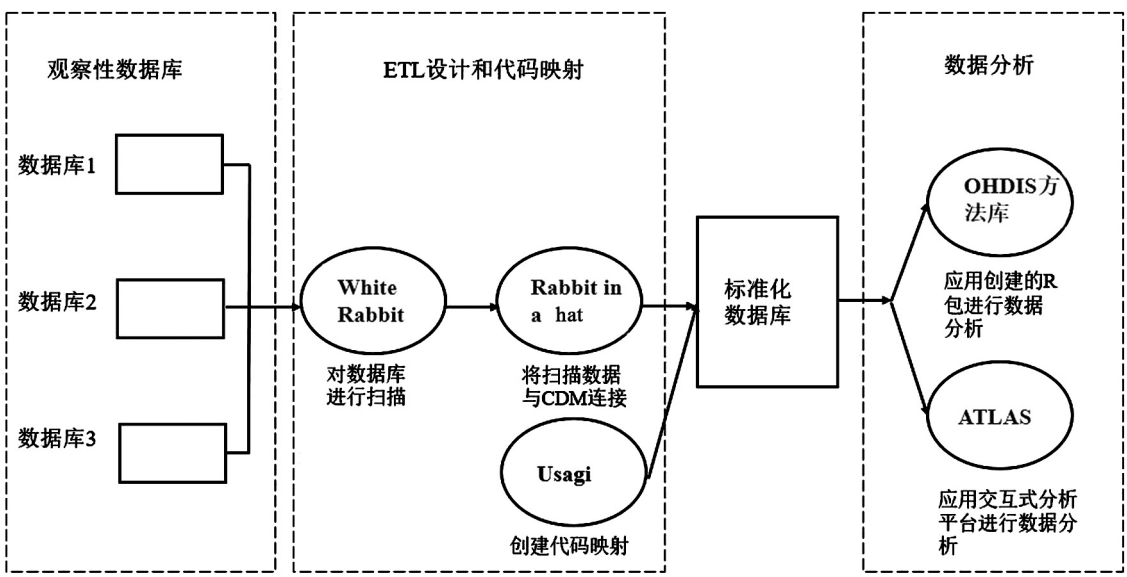

图 1　OHDSI 组织开源工具软件工作流程图

5 类降压药物的疗效与安全性。以下将对 OHDSI 组织这 5 款开源的工具软件进行简要介绍,希望能为我国的 ADR 主动监测提供新思路、新方法。

2.1　White Rabbit

White Rabbit 是 OHDSI 组织设计用来对源数据库进行扫描并创建一份含有必要信息的表格、字段、字段值的 Excel 表格的一款软件。产生必要信息的表格,是由源数据库转化为 OMOP CDM 的基础。该软件及其相应代码可在 https://github.com/OHDSI/WhiteRabbit 下载。

White Rabbit 的优点之一是不仅可以扫描逗号分隔值文本文件,还同时支持 MySQL、SQL Sever、Microsoft Access 等多种数据格式,充分解决了以往扫描软件无法应对多种格式数据库的缺点。其次,White Rabbit 导出的必要信息表中,有意地防止生成患者的身份信息,可以充分保护患者的隐私,有效防止患者信息泄漏。这也是该软件最大的好处。最后,该软件具有一定的数据分析能力,在生成的 Excel 表格中,源表格中的每列会分为两列,一列将会展示所有非重复的值(distinct values),且该值出现的次数需要大于设定的最小频数,如果小于设定的最小频数,则显示为列表截断(list truncated)即不显示。非重复的值可以显示源数据的每列信息中对应的字段、字段值和值的频率。

2.2　Rabbit in a hat

OHDSI 组织设计 Rabbit in a hat 软件的目的是将已经通过 White Rabbit 扫描的源数据所生成的必要信息表与 OMOP CDM 中各个表格和列进行连接。通常 ETL 团队会开会先共同决定对表的映射,随后是列到列以及字段对字段的映射。需要注意的是 Rabbit in a hat 为 ETL 过程生成 Word 文档,但不生成创建 ETL 的代码。该软件及其相应代码可在 https://github.com/OHDSI/WhiteRabbit 下载。

在各个表格映射的过程中,OHDSI 组织建议首先映射 PERSON 表,因为 CDM 是以个人为中心,将不同维度的多个表格连接在一起。随后映射 OBSERVATION_PERIOD 表,因为数据库中每位患者的资料中都会有这个表格,且该表格包含了大部分事件的信息,所以一般在 PERSON 表后进行映射。随后将其他维度的表格与 CDM 的表格进行意义映射。其中,VISIT_OCCURRENCE 是整个映射过程中最复杂的一个表格,因为就诊过程中包含的信息量较大。所以在该表映射过程中,可以运用中间表,但映射完成后,OHDSI 不建议继续应用中间表。

2.3　Usagi

Usagi 是 OHDSI 组织开发用来辅助手动创建代码映射的软件。首先将源数据导入,如果源数据不是英语,一般会通过谷歌翻译将专业术语转为英语,并创建翻译列。运用术语相似性的方法,将源数据或翻译列的数据与 CDM 中的标准概念进行映射,最后需要人工手动进行评审,该过程推荐选择具有医学术语和编码系统经验的专业人士进行审核。最后,将映射的表格导出到 SOURCE_TO_CONCEPT_MAP 中。该软件及其相应代码可在 https://github.com/OHDSI/Usagi 下载。

Usagi 在一般情况下,是将源数据映射到 CDM 的标准概念中。对源数据进行匹配时,会出现“CONCEPT_ID”列,该列将显示匹配分数(匹配分数由 0 到 1,其中 1 为可信匹配,0 为不可信匹配)。如果匹配分数在 0~1 之间,则需要专业人员对匹配进行最终确定,如果认为匹配成功,则选择确认匹配。

Usagi 还设有一些特殊的过滤器选项,如“过滤标准概念”选项,当关闭该选项时,Usagi 还可以考虑映射到 CDM 中的分类概念。其中“通过自动选择的概念/ATC 代码过滤”的过滤器选项较为特殊。当存在限制搜索信息时,对“自动概念 ID”提供“ATC”列表进行限制。即使 ATC 代码不能与标准概念一一对应,但是也将有效地控制搜索空间,提高效率。

2.4　OHDSI 方法库

OHDSI 创建了 OHDSI 方法库(OHDSI Methods 库),即 R 包,以方便研究者们快速调用,避免从头开始编写代码,提高研究效率。OHDSI 方法库主要包含了 3 个部分即预测和估计方法(prediction and estimation methods)、方法描述(method characterization)和支持类 R 包(supporting packages)。这些 R 包为实现完整的观察性研究提供了多种函数,涵盖了从 CDM 数据的估计、统计,以及汇表制图等全过程。基于 CDM 数据,这些 R 包不仅可以提供基础分析而且还可以根据需要提供高级的标准化分析,如人口特征分析、人群水平估计和患者水平预测等。

预测和估计方法部分主要包含了 Cohort Method、Self-controlled Case Series、Self-controlled Cohort、Patient Level Prediction、Case-control、Case-crossover 等 R 包,这些 R 包可以提供高阶的基于 CDM 数据的统计分析。方法描述部分包含 Empirical Calibration、Method Evaluation、Evidence Synthesis 等 R 包。

支持类 R 包部分包含 Database Connector、Sql Render、Cyclops、ParalleLogger、Feature Extraction 等 R 包,其中 Database Connector 包和 Sql Render 包可以连接到各种数据库平台(如 PostgreSQL、SQL Server 和 Oracle 等),Cyclops 包基于高性能电脑可以实现一个高效的回归引擎,能够被所有 OHDSI 方法库的 R 包调用,以实现大规模的回归计算(高维变量、大量观测)。

OHDSI 方法库的 R 包经过了验证检验,于传统的用户自行从头开始编程显得尤为高效、准确,并且不易出错。OHDSI 方法库的 R 包均是开源的,研究者可以下载所需 R 包,进行相应的数据分析、图表绘制等项目。

2.5　ATLAS

ATLAS 是 OHDSI 组织开发的一个基于 Web 界面的交互式分析平台,可以更便于研究人员进行研究。ATLAS 与 OHDSI WebAPI 一起作为 Web 应用程序部署,通常两者一起安装,在《ATLAS GitHub 存储库设置指南》[10] 和《WebAPI GitHub 存储库安装指南》[11] 中可获取每个组件的安装指南。有一个公共的 ATLAS 可以访问一些模拟小型数据,达到练习的目的。这个工具平台使不善于编程的研究人员可以高效、快速地进行数据分析,并且可以生成相应执行过程的 R 语言代码,该代码可以在无安装 ATLAS 与 OHDSI WebAPI 的具有 CDM 的环境下运行,为研究者们提供了极大的便利。

ATLAS 平台提供了 12 个功能模块,分别为数据源(data sources)、词汇搜索(vocabulary search)、概念集(concept sets)、队列定义(cohort definitions)、队列路径(cohort pathways)、发病率(incidence rates)、个人档案(profiles)、人群水平的估计(population level estimation)、患者水平的预测(patient level prediction)、作业(jobs)、配置(configuration)、反馈(feedback)。研究者可以根据不同模块实现不同的功能,如“数据源”的主要功能有查看每个数据源的描述性报告;“词汇搜索”项具备检索 OMOP 标准化术语的能力;“配置”项具有查看已配置数据源的功能等。

ATLAS 平台可以调用 OHDSI 方法库 R 包来实现数据分析的功能(如发病率、人群水平的估计,以及患者水平的预测等),但并不支持所有的方法,如果研究者有特殊的数据分析需求,可以使用 OHDSI 方法库 R 包或是自己通过 SAS 或 R 等软件自行编程进行更加灵活、高效的数据分析。

3　结语

我国以往对上市后药品的 ADR 监测主要是基于自发呈报系统的被动监测为主,虽然具备快速监测信号、花费小、数据多等优势,但其存在着漏报严重、重复上报、不易计算发生率等缺点[12]。主动监测可以很好地弥补被动监测,并且具有计算 ADR 发生率等优势,可以显著提升我国 ADR 监测的整体水平,更好地保障我国人民用药安全。我国 ADR 主动监测起步虽然比较晚,但是进展快速。2017 年 2 月,在国务院发布的《“十三五”国家药品安全规划》中,明确提出要利用医疗机构电子数据,建立药品医疗器械安全性主动监测与评价系统,这表明 ADR 主动监测已成为我国卫生医疗系统的重要工作之一。国家药品监督管理局相关部门组织学术界、制药工业界以及相关机构代表等组成课题组依据部分国外药品监管机构对如何利用真实世界证据支持监管决策的相关指导原则或框架文件反复研讨,旨在进一步规范相关工作,促进药物研发工作质量和效率的提升。于 2018 年 8 月在第八届中国肿瘤学临试验发展论坛上发布我国的首个真实世界研究的指南——《2018 年中国真实世界研究指南》;2019 年 5 月 29 日,国家药品监督管理局药品审评中心发布《真实世界证据支持药物研发的基本考虑》(征求意见稿);2019 年 6 月,国家药品监督管理局与海南省联合启动海南临床真实世界数据应用试点,探索将临床真实世界数据用于药品医疗器械产品注册和监管决策实践,同年 12 月,国家药品监督管理局医疗器械技术审评中心发布《真实世界数据用于医疗器械临床评价技术指导原则(征求意见稿)》,2020 年 1 月 7 日发布国内首个《真实世界证据支持药物研发与审评的指导原则(试行)》;2020 年 8 月国家药品监督管理局药品审评中心发布《真实世界研究支持儿童药物研发与审评的技术指导原则(试行)》,这些文件的发布,充分显示了真实世界数据在我国备受重视,同时如何利用真实世界数据,尤其是将多种机构之间进行数据整合、统计分析,显得尤为重要。

我国目前主要的医疗数据主要来源于医院信息系统(hospital informational system, HIS)、电子医疗记录(electronic medical record, EMR)[13] 等。2016 年由国家药品不良反应监测中心发起建设中国医院药物警戒系统(Chinese hospital pharmacovigilance system, CHPS)[14],其目的在于借助该系统将各数据

(下转第 846 页)

Meta 分析[J]. 天津中医药, 2021, 38(3): 350-356
24 于辉, 汪月奔, 甄军海, 等. 参附注射液对脓毒症患者临床疗效的 Meta 分析[J]. 中国中医急症, 2019, 28(1): 29-33
25 李娜, 蒋林伟, 俞璐, 等. 血必净注射液治疗脓毒症的系统评价[J]. 中国现代药物应用, 2013, 7(22): 8-11
26 哈雁翔, 王晓鹏, 黄坡, 等. 生脉注射液治疗脓毒症休克效果的系统评价和 Meta 分析[J]. 中国中医急症, 2019, 28(11): 1893-1898,1915
27 中华医学会重症医学分会. 中国严重脓毒症/脓毒性休克治疗指南(2014)[J]. 全科医学临床与教育, 2015, 13(4): 365-367
28 陆黎黎. 中医药治疗临床期糖尿病肾病的 meta 分析[D]. 北京:北京中医药大学硕士学位论文, 2019
29 胡晶, 商洪才, 李晶, 等. 血必净注射液治疗脓毒症的系统评价[J]. 解放军医学杂志, 2010, 35(1): 9-12
30 蒋华, 庄燕, 王醒, 等. 活血化瘀法治疗脓毒症的系统评价[J]. 中国中医急症, 2014, 23(12): 2161-2163, 2176
31 高敬书, 王桂媛, 刘松江, 等. 中医临床试验数据收集的短板与对策[J]. 中华中医药杂志, 2016, 31(10): 3881-3883
32 国家药品监督管理局. 国家药监局关于促进中药传承创新发展的实施意见[Z]. 2020
33 肖小河, 柏兆方, 王伽伯, 等. 中药安全性评价与药物警戒[J]. 科学通报, 2021, 66(Z1): 407-414

(2021-08-16 收稿　2021-09-21 修回)

(上接第 840 页)

源的数据通道打通,方便研究者开展安全性监测、评价等研究工作。CHPS 系统内置了通用数据模型,但是只是简单地对数据进行整合,尚不能支持提供进行临床研究级别的整合数据。我国的 CHPS 系统可以借鉴 OHDSI 组织的相关工具软件,并根据我国数据的具体情况对上述工具软件进行进一步完善,以更好地服务于我国的 ADR 主动监测,更好地保障人民用药安全。

参 考 文 献

1 郭晓晶, 王蒙, 郭威, 等. 药品不良反应主动监测中混杂因素控制的现状及挑战[J]. 中国药物警戒,2018, 15(10): 595-599
2 王丹. 药品不良反应主动监测及其发展趋势[J]. 中国药物警戒, 2015, 12(10): 600-602,610
3 宋佳芳, 朱贺, 韩晟. OHDSI/OMOP CDM 在药品不良反应监测中的应用[J]. 医药导报, 2019, 38(1): 54-58
4 何家双, 肖晓旦. OMOP CDM 在临床科研中的应用思考[J]. 中国数字医学, 2016, 11(3): 72-74
5 Observational Health Data Sciences and Informatics. Who we are[EB/OL]. (2021)[2021-05-13]. https://ohdsi.org
6 王玲. 美国观察医疗结果合作项目中数据组织及通用数据模型的应用研究[J]. 中国药物警戒, 2015, 12(6): 341-346
7 Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: Rationale and design for the Observational Medical Outcomes Partnership[J]. Ann Intern Medicine, 2010, 153(9): 600-606
8 Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research[J]. J Am Med Inform Assoc, 2012, 19(1): 54-60
9 Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis[J]. Lancet, 2019, 394(10211): 1816-1826
10 OHDSI. ATLAS setup guide[EB/OL]. (2021)[2021-08-10]. http://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide
11 OHDSI. WebAPI installation guide[EB/OL]. (2021)[2021-08-10]. http://github.com/OHDSI/WebAPI/wiki/WebAPI-Installation-Guide
12 侯永芳, 宋海波, 刘红亮, 等. 基于中国医院药物警戒系统开展主动监测的实践与探讨[J]. 中国药物警戒, 2019, 16(4): 212-214
13 李洪, 魏来, 贾继东, 等. 观察性临床研究是随机对照临床研究的重要补充[J]. 中华肝脏病杂志, 2015, 23(5): 389-392
14 侯永芳, 宋海波, 刘红亮, 等. 基于中国医院药物警戒系统开展主动监测的实践与探讨[J]. 中国药物警戒, 2019, 16(4): 212-214

(2021-06-25 收稿　2021-08-21 修回)