山东农业大学学报(自然科学版),2019,50(3):434-437 VOL.50 NO.3 2019
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2019.03.016

基于 OMOP 通用数据模型的 FAERS 数据库标准化与数据挖掘

张正宇[1],于 跃[2],周 虎[1],赵文龙[1*]

1. 重庆医科大学医学信息学院, 重庆 400016
2. 美国梅奥医院数字医学科学系, 明尼苏达州 55901

摘 要: 应用 OMOP 通用数据模型，对 FAERS 数据库进行标准化转化，通过标准化前后数据质量与数据采集速度的对比分析，展示 OMOP CDM 在 FAERS 数据标准化程中的重要意义。然后，标准化的基础上，对 5-羟色胺再摄取抑制剂（Selective Serotonin Reuptake Inhibitor，SSRIs）药物不良反应信号进行了挖掘，展示了基于“真实世界数据”的 SSRIs 上市后的安全信号的综合挖掘结果，为后续用药研究提供参考。

关键词: OMOP 通用数据模型; 数据标准化; 数据挖掘; 药品不良反应

中图法分类号: TP274 **文献标识码**: A **文章编号**: 1000-2324(2019)03-0434-04

Standardization and Data Mining of FAERS Database Based on OMOP Common Data Model

ZHANG Zheng-yu[1], YU Yue[2], ZHOU Hu[1], ZHAO Wen-long[1*]

1. *School of Medical Information/Chongqing Medical University, Chongqing* 400016, *China*
2. *Department of Digital Medical Science, Mayo Clinic, Minnesota* 55901, *USA*

Abstract: In this study, we utilize OMOM Common Data Model to standardize FAERS data set. And then we evaluate the transformation results to validate the significance of the FAERS standardization. Then, we implement a data mining research about 5 Selective Serotonin Reuptake Inhibitor (SSRIs) drugs base on the standardized FAERS database. The study based on ADR signals in the real world is helpful to evaluate the post-marking safety drugs and provide references for safety in clinical medication.

Keywords: OMOP universal data model; data standardization; data mining; adverse drug reactions

美国食品药品监督局的不良反应上报系统数据库（FDA Adverse Event Reporting System, FAERS）是世界范围内药物监管部门和学术界最常用的药物不良反应检测数据来源之一。但由于 FAERS 中的不良事件数据来源于自发上报，因此其存在一定程度的数据质量问题。随着电子健康档案（Electronic Health Records, EHR）数据库的发展，使得应用 EHR 的“真实世界数据”进行药物不良反应检测与验证成为了可能[1]。而 EHR 与 FAERS 数据的异质性，给药物不良反应挖掘分析带来了困难。因此，为了提高药物不良反应信号挖掘的准确性，对并且为未来 FAERS 与 EHR 相结合进行数据挖掘提供统一的标准化数据，亟需对 FAERS 进行数据标准化。

健康观测数据科学和信息学组织（Observational Health Data Sciences and Informatics，OHDSI）开发的观察医疗结果合作项目通用数据模型（Observational Medical Outcomes Partnership Common Data Model, OMOP CDM）为 FAERS 数据库的标准化和整合提供了框架[2]。OMOP CDM 是一个为医学数据标准化而设计的数据模型，其基本思想是通过统一的数据模型与医学概念词汇表示，使得不同来源的医学数据以统一的标准进行整合。

本课题组的于跃等[3]开发了数据库转化工具 ADEpedia-on-OHDSI，该工具具有较高的数据转化率，可以将 FAERS 数据库较为完整的转化为 OMOP CDM 格式。本文在基于 OMOP CDM 对 FAERS 数据库进行标准化的基础上，对 5-羟色胺再摄取抑制剂（Selective Serotonin Reuptake Inhibitor，SSRIs）药物不良反应信号进行了挖掘。通过标准化前后数据质量的对比分析，展示 OMOP CDM 在 FAERS 数据标准化与挖掘过程中的重要意义。

1 FAERS 数据标准化与挖掘方法

收稿日期: 2018-03-05 **修回日期**: 2018-05-06
基金项目: 基于临床大数据的医疗行为分析系统研究与开发(cstc2015shmszx10004)
作者简介: 张正宇(1994-),女,硕士研究生,主要研究方向为数据挖掘和医学信息. E-mail:389136875@qq.com
***通讯作者**: Author for correspondence. E-mail:cqzhaowl@163.com

数字优先出版:2019-05-24 http://www.cnki.net

1.1　数据来源

数据来源于美国食品与药品监督管理局（Food and Drug Administration，FDA）建立的药品不良事件（Adverse Drug Event，ADE）上报系统(FDA Adverse Event Reporting System, FAERS)数据库[4]。

在不良反应挖掘研究对象的选择上，选取临床广泛使用的 SSRIs 类抗抑郁药物。选择目前常用的五种 SSRIs 类药物：共五种：氟西汀（Fluoxetine）、帕罗西汀（Paroxetine）、舍曲林（Sertraline）、氟伏沙明（Fluvoxamine）以及西酞普兰（Citalopram）作为不良反应挖掘的实验对象，并纳入了 2013 年 1 月 1 日-2017 年 12 月 31 日的药品不良反应（Adverse drug reaction，ADR）信号进行检测。

1.2　基于 OMOP CDM 的 FAERS 数据标准化与数据挖掘框架

设计基于 OMOP CDM 的 FAERS 数据标准化与数据挖掘框架。整个框架主要分为三部分，FAERS 数据标准化、标准化药物不良反应数据查询与提取，基于标准化数据的药物不良反应挖掘。

1.3　数据标准化

采用 OHDSI 组织开发的 OMOP 通用数据模型完成 FAERS 数据库的标准化工作。OMOP CDM 的最大特点是除了提供完备统一的标准化数据库结构外，还提供了用于医学概念的标准化医学词汇表。OMOP CDM 的基本结构如图 1 所示目前，OMOP CDM 中共收录了 116 种不同的医学词汇表/本体，并且通过同义词表，为每一个医学数据设定一个标准的概念映射，使不同数据库之间医学概念描述的差异化问题得到了解决。

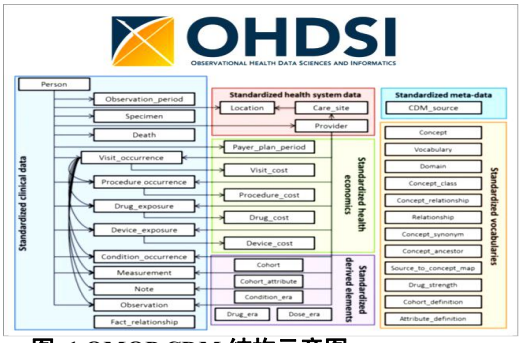

图 1 OMOP CDM 结构示意图

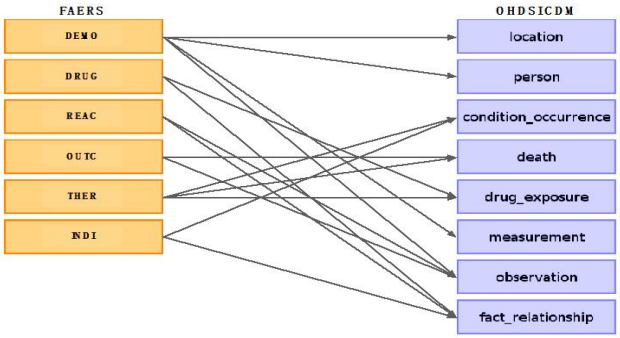

图 2 表级别 FAERS-OMOP CDM 数据结构匹配

Fig.1 OMOP CDM structure schematic　Fig.2 Schematic diagram of faers-omop CDM data structure matching at the table level

在 FAERS 数据库的标准化方面，采用 ADEpedia-on-OHDSI 工具[3]将 FAERS 转化成为了 OMOP CDM 格式。其基本转化流程分为四步。1）数据预处理。应用斯坦福大学 Banda 等人开发的 AEOLUS 工具[5]，完成对 FAERS 原数据中进行数据去重与药物名称标准化等预处理工作。经 AEOLUS 工具处理后 FAERS 数据库中的药物名称被映射到 RxNorm 药物标准化本体[6]，完成药物名称的标准化工作。2）数据结构匹配。将 FAERS 原始的数据结构与 OMOP CDM 的数据框架在逻辑上进行了匹配，用以指导进一步的数据转换工作。主要包括表级别的匹配和字段级别的匹配。表级别的匹配结果如图 2 所示。3）数据提取、转化与加载。在逻辑匹配的基础上，根据 OMOP CDM 的具体要求，对 FAERS 原数据进行了数据的具体转化工作。数据提取、转化与加载内容具体包括：数据类型的转换、医学概念数据的标准化、数据计算、遗失数据插补、数据加载等流程。4）标准化结果评价，为了对数据标准化的结果进行评估，还对数据的转化率、医学概念匹配的正确率、数据计算插补的准确率等进行了评估。以反应整个 FAERS 数据库转换的效果。

1.4　标准化数据查询与提取

应用 OMOP CDM 进行数据标准化与整合的一个重要目的，就是为了实现标准化的数据查询与提取。由于 OMOP CDM 中所有的医学数据均会匹配到标准词汇表中的概念上，因此只要根据标准词汇制定标准化的查询语句，就可以实现不同数据库、甚至不同机构之间的标注化数据查询与提取，既实现了异构数据的标准化查询，又节省了编写查询语句的人力与时间。

数据提取采用根据 OMOP CDM 首选用于标注药物概念的 RxNorm 药物标本体获取 SSRIs 的各具体药物规范名称与概念唯一标识符（RxCUI）与相对应的 OMOP 概念标识符。进而根据编写标准化的 SQL 查询语句完成标准化数据的提取。数据提取完成后还要转置成为“药物-不良反应”矩阵格式，以备接下来的数据挖掘研究使用。

1.5　药物不良反应挖掘方法

基于药品不良反应的数据挖掘方法主要包括比例失衡法（Disproportionality Analysis，DPA）、信息成分法（Information component，IC）、MGPS 相对比值比法（Multi-item gamma passion shrinker，MGPS）、和聚类分析法（clustering or database segmentation）等。其中比例失衡算法包括报告比值比法（Reporting odd radio,ROR）、比例报告比值比法（Proportional reporting radio，PRR）、和贝叶斯置信传播神经网络算法（Bayesian Confidence Propagation Neural Network，BCPNN）等。ROR 法具有较高灵敏度，早期发现 ADR 信号的能力较好，故采用该方法。警戒信号检测标准为：（1）a≥3；（2）ROR 95%CI 下限＞1 提示生成 1 个可疑药物不良反应信号。

MedDRA 不仅用于对药品不良事件的规范化处理和编码，还提供药品不良事件的分类信息。将挖掘出的 ADR 信号按照 MedDRA 的系统器官分类（System organ class，SOC）进行统计整理。MedDRA 所有术语都被赋予唯一的编码，并将其分为系统器官分类、高位组语（High Level Group Term，HLGT）、高位语（High Level Term，HLT）、首选语（PT）和低位语（Lowest Level Term，LLT）5 个层级。基本单元是 PT，用于对医疗事件进行划分和检索。采用 MedDRA19.0 版本对药品不良事件记录在 26 个 SOC 分类上的分布情况进行统计。并应用双聚类算法，绘制不良反应信号的热图，以实现挖掘结果的可视化展示。

2　实验结果及分析

2.1　FAERS 标准化结果

从 FAERS 官网上下载 2013 年 1 月 1 日-2017 年 12 月 31 日的数据进行试验。FAERS 原始数据中共有病人数据 11 904 580 条，经过去重复后，病人数据为 9,956,310 条。进一步对去重复后的数据进行标准化并将其存入 OMOP CDM 数据库中。两个数据库主要表格间的转化结果如表 1 所示。从表 1 可以看出，FAERS 数据库中患者基本数据，临床用药数据，用药适应症数据均全部加载到了 OMOP CDM 相对应的表中。而 FAERS 中的不良反应数据和临床结果数据也被全部转加载到 OMOP CDM 的 OBSERVATION 中（OBSERVATION 表中数据总数等于 FAERS 数据库中 REAC 和 OUTC 两个表数据总数之和）。

另外，本研究同样调查了数据库中医学概念数据标准化的准确率。其中，药物名称匹配成功率约为 94%，仅有 6%左右的药物名称无法被匹配到 OMOP CDM 规定的 RxNorm 标准药物概念上。而不良反应概念与适应症概念由于 FAERS 中已经应用 MedDRA 词表对其进行标注。因此其可以全部转化到 OMOP CDM 的标准概念上。另外，患者的性别、国籍、服药方式、服药剂量等等相关概念的匹配成功率均在 94%以上。说明转化过程中的信息损失较小，不会对后续分析结果造成较大影响。

表 1 FAERS 与 OMOP CDM 数据库标准化前后主要表格数据比较

Table 1 Comparison of main table data before and after the standardization of FAERS and OMOP CDM database

FAERS 表（去重后） Table FAERS after removing duplication	数据量 Data size	OMOP CDM 表 Table OMOP CDM	数据量 Data size
DEMO	9,956,310	PERSON	9,956,310
DRUG	37,288,989	DRUG_EXPOSURE	37,288,989
INDI	23,012,045	CONDITION_OCCURRENCE	23,012,045
REAC	32,504,326	OBSERVATION	40,340,720
OUTC	7,836,394		

2.2　药物不良反应挖掘结果

经ROR法计算得到的五种SSRIs类药物不良反应信号数量如图3所示。其中共有ADR信号187例。

进一步对不良反应信号 MedDRA 术语集进行 SOC 分类，共涉及到 26 个 SOC。绘制热点图（HeatMap）对挖掘出的不良反应信号在人类系统器官级别分类层次进行可视化展示。由图 4 可见，药物不良反应累积的器官/系统主要集中在各类精神类疾病、神经系统疾病、各类检查、胃肠道系统以及血管及淋巴管等系统。

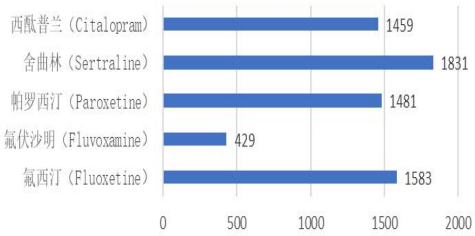

图 3 SSRIs 类药物不良反应信号数量

Fig.3 Number of adverse reaction signals of SSRIs

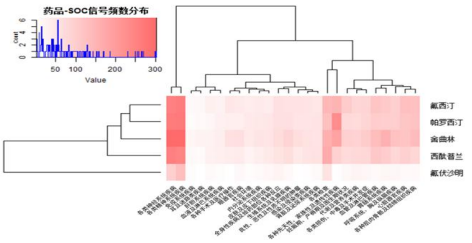

图 4 SSRIs-SOC 分类层次不良反应信号热点图

Fig.4 SSRIs-SOC classification hierarchy of adverse reaction signal heat map

图 4 展示了药品不良反应信号的分布状况。横轴代表 SSRIs 药物的类别，纵轴代表不良事件的 SOC 分类，行与列的交叉处的每个小格代表曲坦类药物和 SOC 分类的组合。每个小格的颜色代表着不良反应信号的频数值，颜色越深，频数越大。白色代表着该“药品-SOC 分类组合”没有探测到药品不良反应信号。

该热点图从 SSRIs 类别和不良事件 SOC 两个维度对不良反应信号进行了聚类分析。首先，从图左侧的聚类树可以看出，主要可以分成两个大类：1）氟西汀、帕罗西汀、舍曲林及西酞普兰涵盖了所有不同层次的不良反应事件，因此将其聚类在相同的类团下；2）氟伏沙明挖掘出的不良反应信号较少，被单独聚到一个类团。相比较氟西汀和帕罗西汀，舍曲林和西酞普兰挖掘出的不良反应信号相对较少，因此该四种药品种又进一步进行划分。从图上侧的聚类树可以看出，5 种 SSRIs 类药物所探测的药物不良反应信号多集中在“各类神经系统疾病”至“各类精神疾病”、“各类检查”至“各种先天性、家族性及遗传性疾病”、“血管及淋巴管类疾病”至“胃肠系统疾病”的 6 个 SOC 分类上，其中氟西汀检测到的危险信号高达 1583 个，氟伏沙明危险信号最少，仅有 429 个。

3　结 论

目前，应用 FAERS 及其它 EHR 进行信号挖掘成为目前药品上市后安全性再评价的研究热点。而数据库中的数据质量问题和不同数据库之间数据异构化的问题是未来药物不良反应检测索要面对的主要困难之一。通过 OMOP CDM 对 FAERS 数据库进行了标准化转化，转化前后的信息损失仅 6% 左右，不会对后续的挖掘分析造成重大影响。基于 OMOP CDM 的 FAERS 数据库标准化优势在于其提高了数据的质量，可以制定可重复使用的标准化查询，提高了数据采集的速度，为未来更多数据库的整合提供了可能。综上所述，本研究为基于“真实世界数据”药物警戒监测工作奠定了基础。

参考文献

[1] Zhou X, Murugesan S, Bhullar H, *et al*. An evaluation of the THIN database in the OMOP common data model for active drug safety surveillance[J]. Drug safety, 2013,36(2):119-134

[2] Hripcsak G, Duke JD, Shah NH, *et al*. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers[J]. Studies in health technology and informatics, 2015,216:574-578

[3] Yu Y, Ruddy KJ, Hong N, *et al*. ADE pedia-on-OHDSI: a next generation pharmacovigilance signal detection platform using the OHDSI common data model[J]. Journal of biomedical informatics, 2019,91:103119

[4] FDA. Questions and answers on FDA's adverse event reporting system (FAERS)[EB/OL]. https://www.fda.gov/drugs/surveillance/fda-adverse-event-reporting-system-faers.html, 2018-01-06/2018-02-06

[5] Banda JM, Evans L, Vanguri RS, *et al*. A curated and standardized adverse drug event resource to accelerate drug safety research[J]. Scientific data, 2016,3:160026

[6] Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio[J]. Pharmacoepidemiology & drug safety, 2004,13(8):519-523