OHDSI

技术指南

实用真实世界临床数据分析工具

前言

这是一本关于观察性健康医疗数据科学与信息学(Observational Health Data Sciences and Informatics)，简称 OHDSI 的书。OHDSI 社区编写此书的目的是希望其能够成为所有 OHDSI 内容的核心知识库，OHDSI 团队会使用开源工具维护并持续更新本书。最新的英文原版可以在 http://book.ohdsi.org 进行免费查看，实体书也可以从亚马逊购买。

本书的目标

本书旨在成为 OHDSI 的核心知识库，主要描述了 OHDSI 小区、OHDSI 数据标准和 OHDSI 软件工具。它对于新手和有经验的使用者都同样适用。它具有很强的实用性，并提供了必要的理论基础及如何操作的一系列说明。读完本书之后，你将理解什么是 OHDSI，以及如何加入我们的工作。你将了解什么是通用数据模型(Common Data Model, CDM)和标准术语集(Standard Vocabularies)等概念，并学习如何使用它们来标准化观察性健康医疗数据。同时你将学习到这些数据的三种应用案例：临床表征、群体水平评估及患者水平的预测。你将会了解到 OHDSI 的开源工具是如何帮助实现这些案例的，同时学习如何使用它们。关于数据质量、临床有效性、软件有效性及方法有效性的章节将会解释如何评估所获得证据的质量。最后，你将学习到如何在分布式协作研究网络中使用 OHDSI 工具来完成这些研究。

本书的结构

这本书由五个主要部分组成：

OHDSI 社区

统一数据表示

数据分析

证据质量

OHDSI 研究实例

每个部分都包含许多章节，每个章节都会根据情况按照介绍、理论、实践、总结和练习这样的顺序进行展开。

撰稿人

本书的每个章节都列出了一位或者多位章节负责人，他们负责该章节的撰稿工作。然而，尚有其他人也对本书的完成做出了贡献，这里我们想要向他们表达感谢：

Hamed Abedtash	Mustafa Ascha	Mark Beno
Clair Blacketer	David Blatt	Brian Christian
Gino Cloft	Frank DeFalco	Sara Dempster
Jon Duke	Sergio Eslava	Clark Evans
Thomas Falconer	George Hripcsak	Vojtech Huser
Mark Khayter	Greg Klebanov	Kristin Kostka
Bob Lanese	Wanda Lattimore	Chun Li
David Madigan	Sindhoosha Malay	Harry Menegay
Akihiko Nishimura	Ellen Palmer	Nirav Patil
Jose Posada	Nicole Pratt	Dani Prieto-Alhambra
Christian Reich	Jenna Reps	Peter Rijnbeek
Patrick Ryan	Craig Sachson	Izzy Saridakis
Paola Saroufim	Martijn Schuemie	Sarah Seager
Anthony Sena	Chan Seng You	Sunah Song
Matt Spotnitz	Marc Suchard	Joel Swerdel
Devin Tian	Don Torok	Kees van Bochove
Mui Van Zandt	Erica Voss	Kristin Waite
Mike Warfe	Jamie Weaver	James Wiggins
Andrew Williams	Chan You Seng	

编译人

本书的编译工作由以下志愿者完成，他们负责各章节的翻译、审校等工作。

主审：(以姓氏笔画为序)

弓孟春	吕晖	朱卫国	刘雷
李劲松	欧玉梅	周毅	徐华

译者：(以姓氏笔画为序)

王昌然	王夏琳	王晓雷	王理
户宜	田亚慧	田雨	周天舒
冯育基	皮轶	刘云	刘迷迷
汤步洲	孙凯奕	李沛尧	李欣
李昱熙	李晋	李倩茹	李静
杨启	杨越	吴琛	何侃
余昶	宋新乔	张宁芮	张瑞霖
张麟	陈清财	林青敏	周天舒
侯丽	闻海妮	姜文卿	洪娜
聂晟	徐昊鹏	唐灵逸	黄立奇
赖俊恺	蔡科	管音	

审校：(以姓氏笔画为序)

弓明	王夏琳	王寅光	王心慰
王琼	王复芹	龙思哲	冯育基
朱之星	刘国臻	汤恺宸	阮彤

孙刚	严晓明	严静东	苏在明
李晋	李智凯	杨洋	李莹
吴垚	佟佳仪	余昶	宋雨
宋怡萱	张武军	张庆超	张迎
陈勇	苗盼盼	范计朋	庞龙
孟祥林	陈宇笛	段芮	闻海妮
赵婷婷	桑田	崔舒雅	郭昱江
梁效玮	梁淑雅	曾玉群	戴芳

秘书：（以姓氏笔画为序）

李星雨	杨越	张迎	徐昊鹏
唐灵逸	管音		

软件版本

本书大篇幅地描述了 OHDSI 提供的开源软件，它们也在不断地更新当中。尽管开发者尽力提供持续稳定的版本，但不可避免的是，软件的更新会使得本书的部分内容相对滞后。社区将会持续升级本书的在线版本，实时反映这些改进之处，实体书的更新版本也会在一段时间后发表。作为参考，以下列出了本书所使用的软件版本号：

- ACHILLES: version 1.6.6
- ATLAS: version 2.7.3
- EUNOMIA: version 1.0.0

•方法集软件包：见表 1

表 1: 本书的方法集中所使用的软件包版本。

Package	Version
CaseControl	1.6.0
CaseCrossover	1.1.0
CohortMethod	3.1.0
Cyclops	2.0.2
DatabaseConnector	2.4.1
EmpiricalCalibration	2.0.0
EvidenceSynthesis	0.0.4
FeatureExtraction	2.2.4
MethodEvaluation	1.1.0
ParallelLogger	1.1.0
PatientLevelPrediction	3.0.6
SelfControlledCaseSeries	1.4.0
SelfControlledCohort	1.5.0
SqlRender	1.6.2

许可证

本书是如何编写并更新的

本书英文版使用 RMarkdown 的 bookdown 包编写。

在线版本使用 Travis 持续集成系统，根据 https://github.com/OHDSI/TheBookOfOhdsi 的资源库进行自动重构更新。每隔一段时间会生成一个本书当前状态的快照，并标记为一个版本。这些版本的实体书都可以从亚马逊获取。

致谢

感谢神州医疗科技股份有限公司，IQVIA 艾昆纬企业管理咨询有限公司，OHDSI 中国工作组，上海交通大学，复旦大学在本书翻译过程中提供的支持和帮助。感谢以上单位成员在本书翻译过程中承担的总协调工作，以及翻译、审校、秘书、通读、排版、设计等工作，保证了本书翻译工作的顺利进行。衷心感谢大家为本书做出的贡献。

免责声明

1. 本书的编译工作均由 OHDSI 志愿者自愿参与，鼓励用户对本书的内容进行应用，以完善您的科学研究。

2. 本书公开发布，对所提供内容的完整性、准确性、及时性不作担保，用户因使用本书而造成的任何损失，由用户自行承担。

3. 对因使用(或不能使用)本书而导致之任何直接、间接、特殊、偶然或结果性损失概不承担责任。

4. 在使用过程中产出的问题或存在疑问，鼓励用户直接与 OHDSI 组织进行交流。

目　录

第一章　OHDSI 社区 1
1.1 从数据到证据的旅程 1
1.2 观察性医疗结果合作组织 2
1.3 OHDSI 作为一个开放科学的合作社区 3
1.4 OHDSI 的进展 3
1.5 在 OHDSI 中进行合作 5
1.6 总结 5
参考文献 5
第二章　从哪里开始 8
2.1 加入探索 8
2.2 怎样融入 12
2.3 总结 14
第三章　开放科学 15
3.1 开放科学 15
3.2 行动中的开放科学：研究马拉松 16
3.3 开放的标准 16
3.4 开放的源代码 16
3.5 开放的数据 17
3.6 开放性讨论 17
3.7 OHDSI 和 FAIR 指导原则 17
参考文献 19
第四章 通用数据模型 21
4.1 设计原理 21
4.2 数据模型约定 22
4.3 CDM 标准化表 27
4.4 其他信息 41
4.5 总结 41
4.6 练习题 41
第五章 标准化术语集 43
5.1 为什么需要术语集，为什么要进行标准化 43
5.2 概念 44
5.3 关联性 50
5.4 层级结构 53
5.5 内部参考表 54
5.6 特殊情形 54
5.7 总结 56

5.8 练习.....56
第 6 章　ETL 技术.....57
6.1 简介.....57
6.2 步骤 1：设计 ETL.....57
6.3 步骤 2：创建代码映射.....62
6.4 步骤 3: ETL 实施.....68
6.5 步骤 4: 质量控制.....68
6.6 ETL 约定及 THEMIS.....69
6.7 CDM 及 ETL 维护.....69
6.8 ETL 最终思考.....70
6.9 总结.....14
6.10 练习.....71
第七章　数据分析用例.....73
7.1 特征.....73
7.2 群体水平评估.....74
7.3 患者水平预测.....75
7.4 高血压方面的应用案例.....75
7.5 观察研究的局限性.....76
7.6 总结.....77
7.7 练习.....77
参考文献.....77
第八章　OHDSI 分析工具.....78
8.1 分析方法.....78
8.2 分析策略.....79
8.3 ATLAS.....79
8.4 Methods 库.....81
8.5 部署策略.....86
8.6 总结.....87
参考文献.....87
第九章　SQL 和 R.....89
9.1 SqlRender.....89
9.2 DatabaseConnector.....96
9.3 查询 CDM.....98
9.4 使用词汇进行查询.....101
9.5 QueryLibrary.....101
9.6 设计一个简单的研究.....102
9.7 使用 SQL 和 R 实施研究.....103
9.8 总结.....108

9.9 练习 108
参考文献 108
第十章 队列定义 109
10.1 队列是什么 109
10.2 基于规则的队列定义 110
10.3 概念集 111
10.4 概率队列定义 112
10.5 队列建立的有效性 112
10.6 定义一个高血压队列 113
10.7 使用 ATLAS 建立队列 113
10.8 用 SQL 构建队列 122
10.9 总结 128
10.10 练习 128
参考文献 129
第十一章 特征描述 138
11.1 数据库层级特征描述 138
11.2 队列特征描述 138
11.3 治疗路径 132
11.4 发病率 133
11.5 高血压患者特征 133
11.6 ATLAS 中的数据库 / 特征描述 134
11.7 ATLAS 中的特征描述 136
11.8 R 软件中队列特征的描述 141
11.9 ATLAS 中的队列路径 144
11.10 ATLAS 中的发病率分析 147
11.11 总结 158
11.12 练习 158
参考文献 159
第十二章 人群水平评估 160
12.1 队列研究设计 162
12.2 自身对照队列设计 155
12.3 病例对照设计 156
12.4 病例交叉设计 157
12.5 自身病例对照系列设计 158
12.6 设计一项高血压研究 159
12.7 使用 ATLAS 进行研究 160
12.8 使用 R 执行研究 170
12.9 研究成果 177

12.10 总结 190
12.11 练习题 191
参考文献 182
第 13 章 患者水平预测 194
13.1 预测问题 195
13.2 数据提取 196
13.3 模型拟合 197
13.4 评估预测模型 201
13.5 设计患者水平预测研究 203
13.6 在 ATLAS 中的操作 206
13.7 在 R 中进行研究 203
13.8 结果外推 208
13.9 其他患者水平预测功能 215
13.10 总结 216
13.11 练习 216
参考文献 217
第 14 章 证据质量 230
14.1 可信证据的属性 218
14.2 理解证据质量 219
14.3 沟通证据质量 220
14.4 总结 221
第 15 章 数据质量 234
15.1 数据质量问题的来源 223
15.2 一般数据质量 224
15.3 研究专用检查 227
15.4 ACHILLES 实践 229
15.5 数据质量控制面板（DQD）实践 230
15.6 研究专用质量检查的实践 232
15.7 总结 234
15.8 练习 234
第 16 章 临床有效性 236
16.1 卫生保健数据库的特点 236
16.2 队列验证 236
16.3 源记录验证 239
16.4 PheValuator 240
16.5 证据的泛化 249
16.6 总结 250
参考文献 250

第 17 章　软件有效性 252
17.1 代码有效性研究 252
17.2 方法库软件开发流程 253
17.3 方法库测试 255
17.4 总结 256
参考文献 256
第十八章 方法有效性（Method Validity） 257
18.1 针对研究设计的诊断方法 257
18.2 可应用于所有估计的诊断方法 258
18.3 实践中的方法验证 263
18.4 OHDSI 方法基准 269
18.5 总结 271
参考文献 271
第 19 章 研究步骤 272
19.1 通用最佳实践指南 273
19.2 详细研究步骤 275
19.3 总结 279
第 20 章 OHDSI 协作网研究 280
20.1 OHDSI 研究协作网 280
20.2 OHDSI 协作网研究 280
20.3 开展 OHDSI 协作网研究 283
20.4 未来展望：协作网研究自动化 285
20.5 最佳的 OHDSI 协作网研究实践 286
20.6 总结 287
附录 288
A 术语表 288
B 队列定义 297
C 阴性对照 313
D 研究方案模版 315
E 参考答案 316

第一章　OHDSI 社区

章节负责人: Patrick Ryan & George Hripcsak

“走到一起是开始，凝聚一心是进步，携手合作才是成功。” -亨利·福特

1.1 从数据到证据的旅程

在世界各地的医疗保健体系中，无论是学术医疗中心，私人诊所，监管机构，医疗产品制造商，保险公司，还是每个病人与医生的交互中，都面临一个共同的挑战：我们如何运用从过去中学到的知识，来为未来做出更好的决策?

十多年来，许多人都在主张建立一个学习型医疗保健体系，“旨在生成并应用最佳证据，协同每个患者和医生一起做出医疗选择；推动在患者护理过程中自然产生、发现证据；并确保卫生服务的创新，质量，安全和价值”(Olsen et al., 2007)。这一雄心壮志的主要部分在于一个令人兴奋的前景 - 即可以对在常规临床护理过程中获得的患者数据进行分析，以产生真实世界的证据，进而可以在整个医疗保健系统中进行传播并影响临床实践。2007 年，美国医学会循证医学圆桌会议发布了一份报告，该报告确立了一个目标：“到 2020 年，90%的临床决策将得到准确，及时和最新的临床信息的支持，并且反映出最佳的可用证据(Olsen et al., 2007)。尽管我们在许多不同方面都取得了巨大进步，但仍然远远没有实现这些值得称赞的愿望。

为什么会这样呢？在某种程度上，因为从患者水平的数据到可靠的证据的过程是艰辛的。没有从数据到证据的单一明确路径，也没有单一地图可以帮助一路导航。实际上，既没有单一的“数据”概念，也没有单一的“证据”概念。

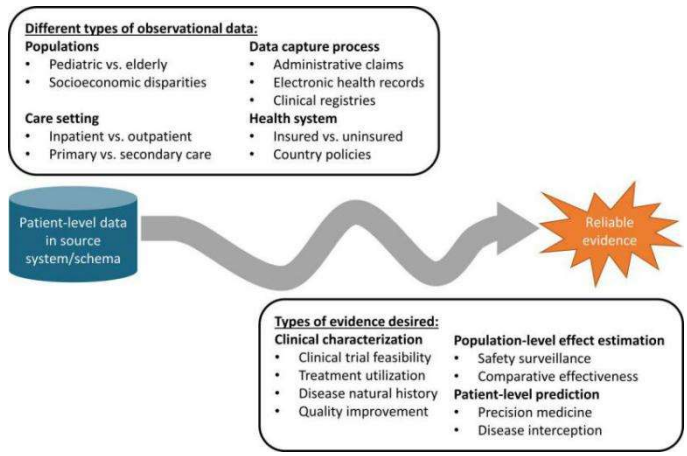


图 1.1: 从数据到证据的旅程

我们现在有不同类型的观察性数据库，它们可以在源系统中获得不同的患者级别的数据。这些数据库与医疗保健系统本身一样多样，反映了不同的人群，医疗机构和数据获得过程。可以利用临床特征、人群水平的效果估计和患者水平的预测的分析用例对不同的证据进行分类，然后来指导决策的制定。除了起点 (源数据)和期望的目的地 (证据)之外，由于进行此过程需要具备广泛的临床，科学和技术能力，因此挑战也进一步复杂化。它需要对健康信息学有透彻的了解，包括从患者和医疗提供者之间的互动，

通过管理和临床系统到最终存储库的全程数据来源，并理解数据采集和管理过程相关的健康政策和行为动机中可能出现的偏差。它需要掌握流行病学原理和统计方法，才能将临床问题转化为观察性研究，来产生相关的答案。它需要具备一定的技术能力，才能对包含数百万例患者的数据集实施和执行有效的数据科学算法，这些数据集需要进行数年的纵向随访。它需要临床知识来整合观察性数据网络中获取的证据与其他信息来源获得的证据，并确定这种新知识应如何影响健康政策和临床实践。因此，很少有人会拥有所有必要的技能和资源来成功地独自完成从数据到证据的过程。取而代之的是，此过程通常需要多个个人和组织之间的协作，以确保使用最适当的方法来分析最佳的可用数据，从而提供所有相关方在决策过程中可以信任和使用的证据。

1.2 观察性医疗结果合作组织

观察性研究合作的一个显著例子是观察性医疗结果合作组织(OMOP - Observational Medical Outcomes Partnership)。 OMOP 是一种公私合作关系，由美国食品药品管理局(FDA - US Food and Drug Administration)主持，美国国立卫生研究院基金会(National Institutes of Health)管理，制药公司联合资助，这些公司与学术研究人员和健康数据合作伙伴合作建立了一项研究计划，旨在利用观察性医疗数据，推动主动医疗产品安全监察科学的发展(Stang et al.，2010)。OMOP 建立了一个多方利益相关者的治理结构，并设计了一系列方法学实验，以对不同的流行病学设计和统计方法的性能进行经验测试，将其应用于一系列保险理赔和电子健康记录数据库，以识别真正的药物安全相关信号，并从假阳性结果中区分出他们。

OMOP 团队意识到，在中心化环境和分布式研究网络中，横跨不同观测数据库进行研究是一项技术挑战，继而他们设计了 OMOP 通用数据模型(CDM)，作为一种标准化机制来规范观察性数据的结构、内容和语义。这样就可以只编写一次统计分析代码，然后该代码即可在各个数据站点重复使用。(Overhage et al., 2012) OMOP 实验表明，建立一个通用的数据模型和标准化术语表是可行的——这一模型可以容纳来自不同医疗健康环境的、用不同来源的术语代表的不同数据类型，实现跨机构协作和高效率计算分析。

从一开始，OMOP 就采用开放型科学的方法，将其所有的工作产出(包括研究设计，数据标准，分析代码和经验结果)开放给公众，以提高透明度，为 OMOP 正在进行的研究建立信心，还提供了一个社区资源来帮助推进他人研究目标。 尽管 OMOP 最初的重点是药品安全，但 OMOP CDM 正在不断发展从而支持更多的分析用例——包括医疗干预措施和卫生系统政策的相对有效性。

尽管 OMOP 成功完成了大规模的实证实验，但(Ryan et al., 2012, 2013b)仍在开发方法上进行了创新， (Schuemie et al., 2014)产生了有价值的知识，这些知识为合理使用观测数据进行安全性决策提供了依据(Madigan et al., 2013b,a；Madigan, Ryan, and Schuemie) 。OMOP 的遗产因为它早日采用的开放式科学原则而被人们铭记，同时它还刺激了 OHDSI 社区形成。

OMOP 项目结束后，它已完成了进行方法学研究以告知 FDA 主动监测活动的任务，该团队意识到 OMOP 旅程的结束将成为新旅程的开始。尽管 OMOP 的方法学研究提供了对科学最佳实践的切实见解，可以明显提高由观察数据产生的证据的质量，但是这些最佳实践的采用却很缓慢。以下是此过程可能遇到的障碍，包括：1)人们对观察数据质量的担忧被认为是分析方法创新之前应优先解决的问题；2)对方法的问题和解决方案的概念理解不足； 3)无法在其本地环境中独立实施解决方案； 4)这些方法是否适用于他们感兴趣的临床问题尚不确定。跨越这些障碍的共同点是，一个人没有办法拥有独自改变

所需的一切能力，但如果有了一些合作的支持，所有问题都可以克服。以下是几个需要合作的领域：

合作建立开放社区的数据标准，标准化术语表和 ETL(Extract-Transform-Load)公约，这将增加基础数据质量的可信度，并促进结构，内容和语义的一致性，以实现标准化分析。

在药物安全性以外的方法学研究上进行合作，以更广泛地建立最佳实践，以进行临床表征(clinical characterization)，人群水平效果评估(population-level effect estimation)和患者水平预测(patient-level prediction)。 在开源分析开发层面进行合作，以整理通过方法论研究证明的科学最佳实践，并作为公开可用的工具，供研究团队轻松使用。共同经历从数据到证据的过程，就临床应用展开合作，以解决整个社区共同关心的重要健康问题。根据这一理念，OHDSI 诞生了。

1.3 OHDSI 作为一个开放科学的合作社区

观察性健康医疗数据科学与信息学(OHDSI，发音为“奥德赛”)是一个开放科学社区，旨在通过各个组织间的合作来收集和分析数据，进而促进更好的决策和医疗。(Hripcsak et al., 2015) OHDSI 进行方法学研究，通过适当地使用观察健康数据，来建立科学的最佳实践，通过开发开源分析软件，将这些实践编码为一致、透明、可重现的解决方案，并将这些工具和实践应用于临床问题，以产生可以指导医疗保健政策和患者护理的证据。

1.3.1 我们的使命

通过各个组织间的合作来收集和分析数据结果，进而促进更好的决策和医疗。

1.3.2 我们的愿景

建立一个通过观察性研究对健康和疾病全面了解的世界。

1.3.3 我们的目标

创新：观察性研究是一个可从颠覆性思维中受益的领域。 我们在工作中积极寻求和鼓励产生新的方法论。

可重复：准确，可重复且经过良好校准的证据对于改善健康至关重要。

社区：OHDSI 欢迎每个人的积极参与，无论是患者，医疗专业人员，研究人员，还是单纯相信我们事业的人。

协作：我们共同努力，优先考虑和解决社区参与者的现实需求。

公开性：我们致力于使我们社区的所有成果都开放并且可以被公开获取，包括我们使用的方法、工具和证据。

公益：我们始终致力于保护社区中每个人和组织的权利。

1.4 OHDSI 的进展

自 2014 年成立以来，OHDSI 不断发展壮大，在其线上论坛吸引了来自学术界，医疗产品行业，监管机构，政府，保险公司，技术提供商，卫生系统，临床医生，患者和不同学科的不同相关方的 2500 多名合作者 ，包括计算机科学，流行病学，统计学，生物医学信息学，健康政策和临床科学。OHDSI

网站上提供了 OHDSI 合作者的列表。[1] OHDSI 合作者地图（图 1.2）突出显示了 OHDSI 跨国社区的广泛性和多样性。

图 1.2: 截至 2019 年 8 月的 OHDSI 合作者地图

截至 2019 年 8 月，OHDSI 还建立了一个数据网络，其中包含来自 20 多个国家的 100 多个不同的医疗数据库，通过使用其维护的开放社区数据标准 OMOP CDM，采用分布式网络方法，收集了十亿多名患者的记录。分布式网络意味着不需要在个人或组织之间共享患者层面的数据。取而代之的是，社区中的个人以研究方案的形式提出研究问题，并伴随分析代码，该分析代码生成的证据是一组汇总的摘要统计数据，并且这些摘要统计数据只在选择进行合作的机构之间共享。借助 OHDSI 分布式网络，每个数据合作伙伴在使用其患者层面的数据时都拥有完全的自主控制权，并遵守各自机构内的数据治理政策。

OHDSI 的开发者们在 OMOP CDM 上创建了一个强大的开源分析工具库，以支持 3 种科研应用：1)临床特征：描述疾病的自然史，治疗利用率和治疗质量的改善； 2)人群水平效果评估：将因果推理方法应用于医疗产品安全性监视和有效性比较； 3)患者水平预测：将机器学习算法应用于精准医学和疾病干预。OHDSI 开发人员还开发了应用程序，以支持 OMOP CDM 的应用，数据质量评估以及促进基于 OHDSI 网络的研究。 这些工具包括 R 和 Python 内置的后端统计软件包，以及 HTML 和 JavaScript 开发的前端 Web 应用程序。所有 OHDSI 工具都是开源的，可通过 Github 公开获得 [2]。

OHDSI 的开放科学社区方法及其开放源代码工具在观察性研究方面取得了巨大进步。最早的基于 OHDSI 网络的研究中，有一个研究分析了横跨三种慢性疾病的治疗途径：糖尿病，抑郁症和高血压。它发表在《美国国家科学院院刊》(PNAS)上，是有史以来规模最大的观察性研究之一，其 11 个数据源的结果涵盖了 2.5 亿多名患者，并且揭示了以前从未观察到的，在治疗选择方面的巨大地理差异和患者异质性（Hripcsak et al., 2016）。OHDSI 已经开发了新的统计方法，用于混杂调整（Tian et al., 2018）和评估因果推断的观察证据的有效性（Schuemie et al., 2018），并且这些方法已在多种情况下被应用了：从癫痫病的个体安全性监测问题（Duke et al., 2017）到二线糖尿病药物的相对有效性（Vashisht et al., 2018），再到针对抑郁症治疗的相对安全性的大规模人群效果评估研究（Schuemie et al., 2018）。OHDSI 社区还建立了一个框架，该框架用于如何负责地将机器学习算法应用于观察性医疗数据（Reps et al., 2018），目前也已应用于各个治疗领域（Johnston et al., 2019; Cepeda et al., 2018; Reps et al., 2019）。

1.5 在 OHDSI 中进行合作

既然 OHDSI 是旨在增强协作能力以生成证据的社区，那么成为 OHDSI 合作者意味着什么？如果您是一个坚信 OHDSI 使命的人，并且有兴趣在从数据到证据的整个过程中做出任何贡献，那么 OHDSI 可以成为您的社区。OHDSI 合作者可以是有权访问患者数据的人，他们有兴趣查看用于生成证据的数据。合作者可以是对建立科学最佳实践和评估替代方法感兴趣的方法学家。可以是软件开发人员，他们有兴趣运用其编程技能来创建可供社区其他人使用的工具。可以是临床研究人员，他们认识到了重要的公共卫生问题，并希望通过出版物和其他形式的传播渠道将这些问题的证据提供给更广泛的医疗界。可以是个人或组织，他们相信影响公共健康的共同原因，并希望提供资源以确保社区能够可持续发展并坚持其使命——包括在世界各地举办社区活动和培训课程。无论您的学科背景或利益相关方，OHDSI 都希望成为一个每个人可以为实现共同目标而共同努力的地方，每个人都做出自己的贡献，共同促进医疗健康的发展。如果您有兴趣加入旅程，请查看第 2 章（“从哪里开始”）以开始行动。

1.6 总结

- OHDSI 的使命是通过各个组织间的合作来收集和分析数据结果，进而促进更好的决策和医疗。
- 建立一个通过观察性研究对健康和疾病提供全面了解的世界，这将通过我们创新的、可复制的、社区的、协作的、公开的和公益的目标来实现。
- OHDSI 合作者专注于开放社区数据标准，方法学研究，开源分析开发和临床应用，以改善从数据到证据的过程。

参考文献

1. Cepeda, M. S., J. Reps, D. Fife, C. Blacketer, P. Stang, and P. Ryan. 2018. “Finding treatment-resistant depression in real-world data: How a data-driven approach compares with expert-based heuristics.” Depress Anxiety 35 (3): 220–28.
2. Duke, J. D., P. B. Ryan, M. A. Suchard, G. Hripcsak, P. Jin, C. Reich, M. S. Schwalm, et al. 2017. “Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network.” Epilepsia 58 (8): e101–e106.
3. Hripcsak, George, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, et al. 2015. “Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers.” Studies in Health Technology and Informatics 216: 574–78. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815923/.
4. Hripcsak, George, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, et al. 2016. “Characterizing treatment pathways at scale using

the OHDSI network." Proceedings of the National Academy of Sciences 113 (27): 7329–36. https://doi.org/10.1073/pnas.1510502113.

5. Johnston, S. S., J. M. Morton, I. Kalsekar, E. M. Ammann, C. W. Hsiao, and J. Reps. 2019. "Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery." Value Health 22 (5): 580–86.
6. Madigan, D., P. B. Ryan, and M. Schuemie. 2013. "Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies." Ther Adv Drug Saf 4 (2): 53–62.
7. Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. 2013. "Evaluating the impact of database heterogeneity on observational study results." Am. J. Epidemiol. 178 (4): 645–51.
8. Olsen, LeighAnne, Dara Aisner, J Michael McGinnis, and others. 2007. The Learning Healthcare System: Workshop Summary. Natl Academy Pr.Overhage, J. M., P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang. 2012. "Validation of a common data model for active safety surveillance research." J Am Med Inform Assoc 19 (1): 54–60.
9. Reps, J. M., P. R. Rijnbeek, and P. B. Ryan. 2019. "Identifying the DEAD: Development and Validation of a Patient-Level Model to Predict Death Status in Population-Level Claims Data." Drug Saf, May.
10. Reps, J. M., M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek. 2018. "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data." Journal of the American Medical Informatics Association 25 (8): 969–75. https://doi.org/10.1093/jamia/ocy032.
11. Ryan, P. B., D. Madigan, P. E. Stang, J. M. Overhage, J. A. Racoosin, and A. G. Hartzema. 2012. "Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership." Stat Med 31 (30): 4401–15.
12. Ryan, P. B., P. E. Stang, J. M. Overhage, M. A. Suchard, A. G. Hartzema, W. DuMouchel, C. G. Reich, M. J. Schuemie, and D. Madigan. 2013. "A comparison of the empirical performance of methods for a risk identification system." Drug Saf 36 Suppl 1 (October): S143–158.
13. Schuemie, M. 2018. "Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data." Proc. Natl. Acad. Sci. U.S.A. 115 (11): 2571–7.
14. Schuemie, M. J., P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan. 2014. "Interpreting observational studies: why empirical calibration is needed to correct p-values." Stat Med 33 (2): 209–18.
15. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. "Improving reproducibility by using high-throughput observational studies with empirical

calibration." Philos Trans A Math Phys Eng Sci 376 (2128).

16. Stang, P. E., P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock. 2010. "Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership." Ann. Intern. Med. 153 (9): 600–606.
17. Tian, Y., M. J. Schuemie, and M. A. Suchard. 2018. "Evaluating large-scale propensity score performance through real-world and synthetic data experiments." Int J Epidemiol 47 (6): 2005–14.
18. Vashisht, R., K. Jung, A. Schuler, J. M. Banda, R. W. Park, S. Jin, L. Li, et al. 2018. "Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative." JAMA Netw Open 1 (4): e181755.

1. https://www.ohdsi.org/who-we-are/collaborators/
2. https://github.com/OHDSI

第二章　从哪里开始

章节负责人：Hamed Abedtash 和 Kristin Kostka

“千里之行，始于足下。”-老子

OHDSI 社区由来自学术界、产业界和政府团体等不同的利益相关者组成。我们的工作能帮助到很多个人和组织，包括患者、服务提供者、科研工作者、以及医疗保健系统、产业界和政府机构。这些好处来源于医疗数据分析质量的提高和医疗数据的新用途。我们认为，观察性研究是一个因颠覆性思考获益良多的领域，所以我们在工作中积极探索并鼓励采用新颖的方法。

2.1 加入探索

无论您是患者、医疗专业人员、研究人员，或只是单纯对我们的工作感兴趣，我们都欢迎您积极参与到 OHDSI 的工作中来。OHDSI 采取一种包容性的成员模式，加入 OHDSI 的成员无需缴纳会员费，而加入 OHDSI 年度会员名单只需举手之劳。同时，这一成员关系是可以随时终止的。OHDSI 的成员可以在这个组织内有不同程度的贡献，比如参与每周一次的组织电话会议，或推动组织的研究工作，又或者在 OHDSI 工作小组内任职。积极活跃的 OHDSI 成员不一定需要是医疗数据的管理人员。OHDSI 的目标是服务于数据管理人员、科研人员、医疗服务提供人员以及患者，或者客户类群体。成员的基本资料会在 OHDSI 的官方网站上公布并定期更新，而成员队伍则通过团体电话会议、工作组以及区域性的分会来发展壮大。

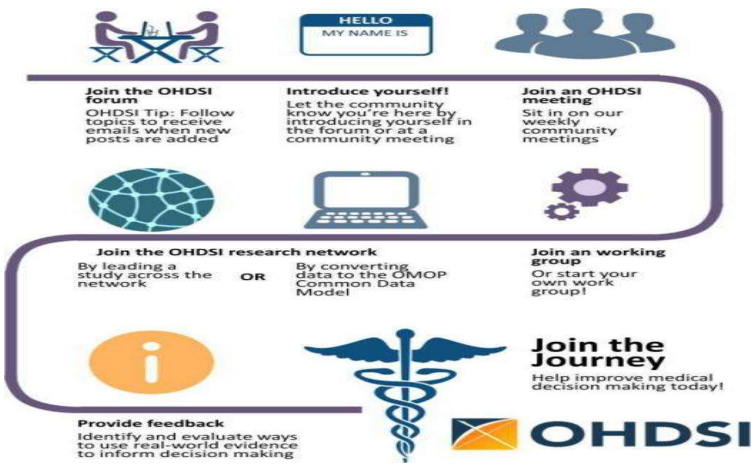


图 2.1: 加入探索旅程- 如何成为一名 OHDSI 合作者

1. 加入 OHDSI 论坛：OHDSI 提示：关注相关主题，以便当新内容发布时，可以收到相关邮件；
2. 自我介绍：让其他成员在论坛或者联盟大会上了解你；
3. 加入一次 OHDSI 会议：参与每周例行的联盟会议；
4. 加入工作组：或者开展自己的工作组！
5. 加入 OHDSI 的研究网络：通过研究网络组织一次研究，或者将数据转化成 OMOP 的 Common Data Model。

6. 提供反馈：确认并评估基于真实世界证据进行决策的方法。

2.1.1 OHDSI 论坛

OHDSI 论坛 3 是一个在线讨论网站，OHDSI 成员可在网站上发布消息，进行对话交流。论坛由树型目录结构组成，最高一级的分类称为“类目”，各论坛可以分成不同“类目”展开相关讨论。“类目”下设有子论坛，每个子论坛又可以进一步延展出下一级的论坛。主题（通常称为线程）位于最低级别的子论坛下面，是论坛成员开始讨论或发布信息的区域。

在 OHDSI 论坛中，你能找到以下类别的内容：

常规：关于 OHDSI 联盟以及如何加入 OHDSI 的一般性讨论；

实施者：探讨如何在当前的环境中实施 Common Data Model 和 OHDSI 分析框架；

开发人员：讨论 OHDSI 应用程序和其他利用 OMOP CDM 工具的开源开发；

研究人员：讨论基于 CDM 的研究，包括证据的建立、协同研究、统计方法和其他有助于 OHDSI 研究网络的主题；

CDM 创建人员：讨论正在进行的 CDM 开发工作，包括需求、术语以及技术方面的工作；

术语集用户：关于术语集内容的讨论；

地区分会（例如韩国，中国，欧洲）：以当地语言进行与当地 OMOP 实施和 OHDSI 联盟活动有关的区域性讨论。

发布自己的主题之前，您需要注册一个帐户。拥有论坛帐户后，您将受邀在常规主题下“欢迎使用 OHDSI！——请做个自我介绍” 线程中介绍自己。请您回复并 1）向我们介绍自己和自己的工作 2）告诉我们您想如何在联盟中做出贡献（例如软件开发、运行研究、撰写研究论文等）。完成了这些，您就进入了 OHDSI 的探索旅程了，欢迎您加入讨论。OHDSI 联盟鼓励您使用这个论坛作为提问、讨论新想法和合作的平台。

你可以选择主题进行“关注”。无论何时，该主题下有新内容发布，你都会受到邮件提醒，并且可以通过回复邮件，直接回复内容的发布者。关注常规主题可以接收即将举办的大会日程和合作机会，同时可以在你的收件箱中收到每周的 OHDSI 文摘!

2.1.2 OHDSI 活动

OHDSI 会定期组织一些线下的活动，为成员提供互相学习的机会，以促进未来更多的合作。这些活动信息会在 OHDSI 官网上发布，只要感兴趣，任何人都可以免费参与。OHDSI 研讨会是学术性的会议，每年在美国、欧洲和亚洲举行，成员们可以通过大会报告、海报展览、软件演示等形式来分享自己的最新研究成果，为促进成员间的交流和了解联盟内最新的成果提供了一个绝佳的分享平台。同时，研讨会通常还会同期开设 OHDSI 课程，由资深 OHDSI 成员担任课程教员，为联盟内的新成员提供亲身参与有关数据标准的主题和对最佳实验分析的机会。这些课程一般是视频录制的，在会议结束后会上传到 OHDSI 的官网上，方便不能到场的成员观看学习。

OHDSI 成员的线下活动是规模较小的讨论会，一般会聚焦某个成员们共同关注的问题，往期活动

包括表型数据编程马拉松（phenotype hack-a-thon）、数据质量编程马拉松（data quality hack-a-thon）、以及开源软件文档马拉松（open-source software documentation-a-thon）等。OHDSI 曾经多次举办连续多天的研究马拉松（Study-a-thon）活动，从而以团队合作的形式，针对特定的研究问题设计和实施合理的观察性分析方法，并通过 OHDSI 研究网络执行研究，形成可以公开发布的证据。在所有这些活动中，我们不仅希望解决共性的问题，同时也要提供一个舒适的环境，来鼓励大家在合作解决问题的过程中不断学习和改进。

如想了解更多 OHDSI 社区的影响力，可访问 OHDSI 网站上“OHDSI 往期活动”部分，探索曾经举办过的座谈会、线下交流活动并观看 OHDSI 教程。“往期活动”内容会定期更新和存档。

2.1.3 OHDSI 社区电话会议

每周一次的 OHDSI 社区电话会议，聚焦于 OHDSI 社区内正在发生的事情。电话会议在美国东部时间每周二下午 12 点-1 点举行，是 OHDSI 成员们在会议中分享最新进展，了解各个成员、工作组以及整个社区研究成果的好机会。每周的会议内容都会被记录下来，会议报告也会存档在 OHDSI 网站资源中。

我们热烈欢迎所有 OHDSI 成员参加每周的电话会议，并希望大家能提出有关社区讨论的主题。可以把 OHDSI 社区电话会议当作一个讨论平台，来共享研究结果，展示进行中的工作情况以寻求反馈，介绍正在开发的开源软件工具，探讨用于数据建模和分析的最佳方法，以及集思广益寻找未来基金申请、研究发表、会议研讨的机会。如果您有一个 OHDSI 会议主题的想法，欢迎在 OHDSI 论坛上发表您的见解。

作为 OHDSI 社区的新成员，我们鼓励您将社区电话会议活动添加到日程中，以了解 OHDSI 研究网络中正在发生的一切。如果您想加入 OHDSI 社区电话会议，请查阅 OHDSI Wiki。社区电话的主题每周都不同。您也可以在 OHDSI 论坛上查阅 OHDSI 的每周摘要，以获取更多关于每周主题的信息。当第一次参加电话会议时，您会被邀请进行自我介绍，借此向社区介绍自己的背景以及加入 OHDSI 的原因。

2.1.4 OHDSI 工作组

OHDSI 工作组团队领导各种正在进行中的项目，每个工作组有自己的领导团队，负责确定项目的短期目标、长期目标。工作组向所有愿意为项目的短期和长期目标而努力的人开放，工作组的设立可以是为了实现长期、战略性的目标，也可以是为实现社区特定需求的短期项目。工作组电话会议频率由项目领导者决定，因工作组而异。活动工作组的列表在 OHDSI Wiki 上可见。

表 2.1 为 OHDSI 活动工作组快速一览表。我们鼓励您参加一次活动，了解更多信息。

工作组名称	目标	目标听众
Atlas 及 WebAPI	Atlas和WebAPI是 OHDSI 开源软件架构的组成部分，目的是提供在OMOP通用数据模型基础上建立标准化分析的能力。	Java和JavaScript软件开发者，有志于改进和为开源Atlas/WebAPI平台作出贡献
CDM 和术语集	进一步开发 OMOP 通用数据模型以满足应用于临床患者数据的系统性、	任何对改进 OMOP 通用数据模型和标准词表

	标准化和大规模的分析软件。通过增加对国际编码系统和临床患者照护的覆盖提高标准词表的质量，以支持其他工作组开发的标准化分析软件。	有兴趣的人，以使之满足所有需求和使用情况
基因组学	扩展 OMOP CDM，与患者基因数据整合，定义CDM适用性框架，可以存储来源于不同测序流程的基因变异信息。	向所有人开放
人群水平评估	开发观察性研究人群水平效果评估的科学方法，具有准确性、可靠性和可复制性，并且帮助社区使用这些方法。	向所有人开放
自然语言处理	以在OHDSI框架下促进使用电子病历（EMRs）中的文本信息为目标，开发应用于临床文本的方法和软件，供OHDSI社区的研究所使用。	向所有人开放
患者水平预测	建立一套开发准确和校正良好的患者为中心的预测模型的标准流程，可用于多种结局和任何患者亚群的观察性格健康数据。	向所有人开放
表型金标准图书馆	使OHDSI社区能够在研究或其他活动中查找、评估和使用经社区验证过的队列定义。	向所有对治疗和表型验证感兴趣的人开放
FHIR 工作组	制定OHDSI和FHIR结合的路线图，为更多学术群体在以下方面提供建议，包括在基于OHDSI的观察性研究中使用FHIR和EHR的数据，通过基于FHIR的工具和API发布OHDSI数据和研究结果。	向所有对互操作感兴趣的人开放
GIS	扩展OMOP CDM并使用OHDSI工具，将患者的环境暴露史和临床表型关联起来	向所有对健康有关的地理归因感兴趣的人开放
临床试验	理解临床试验使用案例，OHDSI平台和生态系统可以在其中任何环节发挥作用，并协助驱动OHDSI工具升级以支持临床试验。	向所有对临床试验感兴趣的人开放
THEMIS	THEMIS的目标是在OMOP CDM规范之上和之外开发标准规范，以确保每	

	个OMOP中心设计的ETL流程是高品质、可复制和高效的。	
元数据和注释	我们的目标是定义通用数据模型中存储人和机器命名的元数据和注释的标准流程，确保研究者能够使用和创造基于观察性研究数据集的数据产物。	向所有人开放
患者产生的健康数据（PGHD）	本工作组的目标是建立PGHD的ETL规范、临床数据整合流程和分析流程，PGHD通过智能手机/App/可穿戴设备产生。	向所有人开放
女性OHDSI	为OHDSI社区内的女性提供论坛，讨论女性在科学、技术、工程和数学（STEM）工作中面临的挑战。我们的目标是为在OHDSI社区如何支持女性在STEM中分享观点、提出担忧、提出建议的讨论提供便利，并最终激发女性成为社区和相关领域的领导者。	向所有认同此项工作的人开放
指导委员会	确保所有OHDSI的活动与正在成长中的社区的需求相一致，以支撑OHDSI任务的愿景和价值。此外，作为哥伦比亚OHDSI协调中心的顾问组，提供OHDSI未来方向的指导意见。	社区内的领导者

2.1.5 OHDSI 区域性分会

OHDSI 区域性分会代表位于同一地理区域的 OHDSI 成员，分会希望通过组织本地网络活动和会议来解决本地的问题。如今，OHDSI 区域性分会包括 OHDSI 欧洲 4、OHDSI 韩国 5 和 OHDSI 中国 6。如果您希望在所在地区成立 OHDSI 区域性分会，您可以遵照 OHDSI 网站上的 OHDSI 区域性分会操作流程进行操作。

2.1.6 OHDSI 协作研究网络

许多 OHDSI 成员希望把研究数据转化到 OMOP 通用模型中。OHDSI 研究网络代表了全球不同背景的观察性研究数据库，这些数据库已经过抽取-转换-加载（ETL）流程从而符合 OMOP 通用模型。如果你的工作涉及到数据转化，那么有许多社区资源都可帮助到你，包括 OMOP 通用模型和术语资料、辅助转换的免费工具和以特定领域或类型数据转换为目的的工作组。我们鼓励 OHDSI 合作者利用 OHDSI 论坛探讨和解决 CDM 转化过程中的难题。

2.2 怎样融入

至此，你可能会有这样的疑惑：我在哪里能融入 OHDSI 社区？

我是临床研究人员，想要开始一项研究任务。如果你是一名临床研究人员，对使用 OHDSI 研究网络来回答特定问题感兴趣，可能甚至想要发表论文，那么你来对地方了。你可以从这里起步，把想法发布到 OHDSI 论坛的 OHDSI 研究者主题下面——这会有助于你和兴趣相同的研究者取得联系。OHDSI 有许多资源可以加快研究问题转化为分析结果，并且乐意发表这些结果和论文。在第 11、12 和 13 章可以找到更多信息。

我想阅读和使用 OHDSI 社区发布的信息。无论你是患者、执业临床医师或是医疗领域的专家，OHDSI 都想要提供高质量的证据以帮助你更好地理解健康结局。也许你已经有阵子没写过代码了，也许你从未学习过编程，但社区里都有你的一席之地。你是证据消费者，也就是将 OHDSI 研究转变为实际行动的人。你正在通过筛查了解 OHDSI 已经或正在产生的证据，可能还想提出与你相关的问题。你开始在 OHDSI 论坛上提问，参加社区电话会议了解最新的研究进展，参加 OHDSI 研讨会和面对面会议来直接参与社区事务。你提出的问题是 OHDSI 社区的重要部分，把问题说出来帮助我们更了解你正在寻找什么证据。

我在医疗领导岗位上工作，可能是数据拥有者和/或代表数据拥有者，正在为所在机构评估 OMOP CDM 和 OHDSI 分析工具。作为机构管理者或领导，你可能听说过 OHDSI，想了解 OMOP CDM 怎么用于你们的研究。你可以从浏览 OHDSI 历史活动材料开始，了解研究的主体部分；可以参加社区会议，简单了解；也可以浏览第 7 章 (数据分析使用案例) 的内容，来帮助你理解 OMOP CDM 和 OHDSI 分析工具能解决的研究类型。OHDSI 社区已为你做好了准备。如果有感兴趣的特定领域，大胆说出来。OHDSI 在世界范围内已有超过 200 个合作机构，丰富的成功案例展现出社区的存在价值。

我是数据库管理者，想要把本机构的数据通过 ETL/转化到 OMOP CDM。选择将数据“OMOP”化是一种创新，也是值得的尝试。如果你刚开始 ETL 流程，请参考 OHDSI 社区 ETL 教程幻灯片或注册下一次在 OHDSI 研讨会上的培训。可以考虑带着问题加入 THEMIS 工作组电话会议和 OHDSI 论坛，你将会发现社区里有很多高手能够帮助你成功完成 OMOP CDM 实施工作。别害羞！

我是生物统计学家和/或方法开发人员，对 OHDSI 工具包的研发感兴趣。你懂 R 语言，了解怎样使用 Git。最重要的是，你愿意把专业技能用于 OHDSI 方法图书馆和进一步的开发。你可以先加入人群水平评估或患者水平预测工作组电话会议，以了解更多目前社区的优先事务。既然你正在使用 OHDSI 工具，你也可以在 GitHub repo 下提问（例如，一个 SQL Render 包问题，你可以在 GitHub Repo 的 OHDSI/SqlRender 下面提问）。我们欢迎您的参与！

我是软件开发人员，对开发可补充到 OHDSI 工具栈中的工具感兴趣。欢迎加入社区！作为 OHDSI 使命的一部分，我们的工具是开源的，受到 Apache 许可证的管理。欢迎你开发可补充到 OHDSI 工具栈的解决方案，请随时加入工作组，抛出你的想法。请记住 OHDSI 在公开科学和公开合作方面投入巨大，我们也欢迎专有算法和软件解决方案，但这不是软件开发的主要努力方向。

我是一名顾问，希望为 OHDSI 社区提供建议。欢迎来到社区！您的专业知识是宝贵的，我们非常重视。欢迎您在适当的时候，在 OHDSI 论坛上推广你的服务。同时作为回报，我们邀请您加入 OHDSI 课程，并考虑在全年的研讨会和 OHDSI 面对面会议上贡献您的专业知识。

我是一名学生，希望了解更多关于 OHDSI 的内容。你来对地方了！考虑加入 OHDSI 电话会议并介绍自己。我们鼓励你钻研 OHDSI 课程，参加 OHDSI 研讨会和面对面会议，学习更多 OHDSI 社区提供的方法和工具。如果你有特定的研究兴趣，可以在 OHDSI 论坛上发布研究主题告诉我们。许多组织提供 OHDSI 资助研究机会 (例如博后、研究奖学金)。OHDSI 论坛将为你提供这些机会的最新消息。

2.3 总结

– 开始 OHDSI 社区之旅就像说“你好”一样容易！在 OHDSI 论坛上发帖并加入社区电话会议。

– 在 OHDSI 论坛上发布你的研究或 ETL 问题。

3. http://forum.ohdsi.org
4. https://www.ohdsi-europe.org/
5. http://forums.ohdsi.org/c/For-collaborators-wishing-to-communicate-in-Korean
6. https://ohdsichina.org/

第三章　开放科学

章节负责人：Kees van Bochove

自 OHDSI 社区成立以来，它的目标就是建立一个基于开放科学价值观的国际合作，例如开源软件的使用、所有会议记录和材料的公开以及新的医学证据透明出版和开放获取。但开放科学到底是什么？医疗数据高度隐私且敏感，通常因这些理由不对外开放，OHDSI 如何围绕医疗数据建立一种开放科学或开放数据策略？为何可重复性的分析如此重要，OHDSI 社区如何实现这一目标？这些是我们在本章中将涉及的问题。

3.1 开放科学

“开放科学”这一概念兴起于上世纪 90 年代。但 20 世纪的第一个十年，即 OHDSI 诞生的这一时期，开放科学才真正获得了关注。维基百科(Wikipedia, 2019a) 将其定义为“一项社会运动，致力于使科学研究（包括出版物、数据、实物样本和软件）及其传播为社会的各个阶层可及，无论是业余还是专业。”开放科学通常通过协作网络逐步完善成熟。尽管 OHDSI 社区从未明确地将自己定位为一个“开放科学”的集体或网络，但该术语通常用于解释 OHDSI 背后的驱动思想和原理。例如，在 2015 年，Jon Duke 将 OHDSI 称为“一种医学证据生成的开放科学方法” 7，2019 年，EHDEN 联盟的前言网络研讨会将 OHDSI 网络方法称为“21 世纪的现实世界开放科学” 8。确实，正如我们将在本章中看到的那样，许多开放科学的实践可以在今天的 OHDSI 社区中找到。有人可能会说，OHDSI 社区是一个“草根”的开放科学团体，其共同愿望是提高医学证据生成的透明度和可靠性。

开放科学或“科学 2.0”（Wikipedia，2019b）方法旨在解决当前科学实践中的诸多已被发现的问题。信息技术导致了爆炸性的数据生成和分析方法的增加。对于个别研究者来说，很难一直跟进其专业领域内发表的所有文献。对于一些将临床实践作为日常工作但仍需要了解最新的医学证据的医生来说更是如此。此外，人们越来越担心许多实验可能存在统计设计缺陷、发表偏倚、P 值操纵和类似的统计问题从而难以重复。纠正这些问题的传统方法是对发表文章进行同行评阅。但这通常无法全部识别和解决上述问题。2018 年《自然》杂志特别版“不可复制研究的挑战” 9 介绍了一些这样的例子。一组试图对其领域文章进行系统性同行评审的作者发现，由于各种原因，他们发现的错误是很难被后续的系统性评价所纠正的，特别是那些在设计上就有严重缺陷的研究。用 Ronald Fisher 的话说：“在完成实验后才去咨询统计学家通常只能要求他对研究进行‘尸检’。他能说出的只有实验为什么失败的原因”（Wikiquote, 2019）。这项工作的研究者发现了一些常见的统计问题，包括随机化设计不佳导致关于统计显著性的错误结论、荟萃分析的错误计算以及不适当的基线比较（Allison et al，2016）。来自同一专题的另一篇文章则以物理学实验为例，提出不仅应该提供基础数据，还要公开和正确记录数据处理过程和分析方法，以确保实验的完全可重现性。这一点至关重要（Chen et al，2018）。

OHDSI 社区用自己的方式应对这些挑战，特别强调大规模产生医学证据的重要性。如 Schuemie 等人所述（2018b），现行的研究范式着重于“使用一个独特的可信度未知的研究设计，每次产生并发表（或不发表）一个评估结果”。但 OHDSI 社区“主张采用一致的和标准化的方法进行高通量观察性研究，实现评估、校准和无偏见的传播，从而生成一个更为可靠和完整的证据库”。 这主要通过将以下步骤整合在一起来实现:构建医疗数据源网络，保证其中的数据映射到 OMOP 通用数据模型;开源共享

分析代码，确保所有人可使用和验证;协调组织大规模基线数据，如发表在 howoften.org 网站上的疾病发生情况信息等。在接下来的段落中，我们将提供具体示例，并以开放标准、开放源代码、开放数据和开放讨论四项原则为指导，进一步详细介绍 OHDSI 的开放科学方法。最后，以开放科学的角度简要介绍 FAIR 原则和 OHDSI 的展望。

3.2 行动中的开放科学：研究马拉松

OHDSI 社区的最新发展是“研究马拉松(study-a-thons)”的出现：即由多学科学者组成简短而集中的面对面会议，旨在使用 OMOP 数据模型和 OHDSI 工具来回答重要的、与临床相关的研究问题。一个很好的例子是 2018 年的牛津大学研究马拉松。该研究在 EHDEN 网络研讨会 10 中得到进一步阐述，研讨会提供了该研究过程的概述，并强调公开可用的结果。在“研究马拉松”正式开始之前，参与者提出医学相关研究问题进行讨论，并在“研究马拉松”过程中选择一个或多个研究问题进行研究。数据是参与者提供的，他们可以访问并且查询 OMOP 格式的患者级的数据。在实际的研究马拉松中，大部分的研究时间用于讨论统计方法（请参见第 2 章）、数据源的适用性、交互产生的结果及这些结果引起的后续问题。就牛津的研究马拉松而言，问题集中在研究不同膝关节置换手术的不良术后反应、在研究马拉松过程中使用 OHDSI 论坛和工具以交互方式发布了研究结果（请参见第 8 章）。OHDSI 工具（如 ATLAS）有助于快速创建、交换、交流、讨论和测试队列的定义，从而极大地加快了就问题定义和方法选择等达成共识的过程。由于所涉及的数据源使用了 OMOP 通用数据模型，且 OHDSI 已完成了开源的患者层面风险预测的研究软件包，这使得我们可以在一天内创建患者术后 90 天死亡率的预测模型，并在第二天从多个大型数据源中对模型进行外部验证。这项研究马拉松还产生了一篇传统的学术论文（全膝关节置换术后不良后果的患者水平预测模型的开发和验证，Ross Williams，Daniel Prieto-Alhambra 等人，正在准备手稿）。该论文耗时数月通过了同行评审。但事实上，在一周之内从零开始完成研究的设想、执行与发布多个医疗数据库的分析脚本与分析结果，覆盖数百万条患者记录，这说明了 OHDSI 可以为医学研究带来根本的改进，将医学证据的处理时间从几个月缩短到几天。

3.3 开放的标准

OHDSI 社区维护的非常重要的资源是 OMOP 通用数据模型（请参见第 4 章）和关联的标准化词汇表（请参见第 5 章）。该模型本身旨在捕获观察性医疗数据，最初旨在分析如药物、手术、设施和结果（如条件和测量结果）之间的联系。现在，OMOP 已扩展至各种案例分析（另请参见 7 章）。协调来自多种编码系统、医疗保健体系和不同类型的医疗保健资源的全球医疗数据，需要在源代码与其最接近的标准版本的代码之间进行大量的“映射”。 OMOP 标准化词汇将在第 7 章中进一步阐述，其中包括来自全球范围内使用的数百种医学编码系统的映射，可以通过 OHDSI Athena 工具进行浏览。这些词汇表和映射是可以免费获得的 OHDSI 社区资源，由此 OHDSI 社区为医疗数据分析做出了重要贡献。从多个方面来看，OMOP 是该领域的最全面的模型，覆盖了约 12 亿条医疗保健记录 11 (Garza et al，2016)。

3.4 开放的源代码

OHDSI 社区提供的另外一个重要的资源是开源的程序源代码。这些程序可被分为几类，比如将数

据映射至 OMOP 的辅助工具（请参见第 6 章）、包含一整套强大的常用统计方法的 OHDSI 方法库、已发表的观察性研究的开源代码以及 ATLAS、Athena 和其他一些支持 OHDSI 生态系统（请参见第 8 章）的基础设施相关的软件。从开放科学的观点来看，最重要的资源之一是一些实际应用过的研究代码，例如那些 OHDSI 研究网络 (请参见第 20 章) 中的研究。这些程序利用了完全开放源代码的 OHDSI 堆栈，可以通过 GitHub 检查、查阅和改进这些堆栈。例如，网络研究通常建立在方法库的基础上，它确保了统计方法在不同分析用例中可被重复使用。在第 17 章更详细概述了 OHDSI 中开源软件的使用和协作如何得以最终保证所生成医学证据的质量和可靠性。

3.5 开放的数据

由于健康医疗数据的隐私敏感的特性，不可能完全开放患者级数据集。但是，可以利用 OMOP 映射的数据集来发布重要的整合分析结果，例如前面提到的 http://howfund.org 网站和发布到 http://data.ohdsi.org 网站的其他公共结果集。此外，OHDSI 社区还提供模拟数据集，如用于测试和开发的 SynPUF。OHDSI 研究网络（请参见第 20 章）可在已将其数据映射到 OMOP 的数据源网络中执行研究。为了使数据源和 OMOP CDM 之间的映射简洁易懂，OHDSI 鼓励数据源机构复用 OHDSI ETL 或“映射”工具，并将其映射代码作为开放源代码发布。

3.6 开放性讨论

开放的标准、开放的源代码和开放的数据是巨大的资产，但它们本身存在不会影响医疗实践。开放科学实践和 OHDSI 所产生的影响取决于其生成的医学证据及基于证据的医学实践。OHDSI 社区在美国、欧洲和亚洲举办了数次 OHDSI 研讨会，并在中国和韩国等地创办了专门的协调组织。这些研讨会讨论了 OHDSI 开源社区在统计方法、数据和软件工具、标准词汇表以及其他方面的进展。OHDSI 论坛 12 和 Wiki 13 为全球数以千计的研究人员提供了观察性研究的便利。在 Github 15 中 OHDSI 社区调用 14 、分发及抽取代码，不断演变出 OHDSI 社区的开放资产，如代码和 CDM。在 OHDSI 网络中，全球性的观察性研究以公开和透明的方式进行，使用全球数亿份患者记录。整个社区都鼓励开放和开放性的讨论。本书即是过程开放的典型例子，作者是通过 OHDSI wiki、社区召集及 GitHub 资源库 16 之间的通力协作完成写作的。需要强调的是，没有了所有 OHDSI 合作者的支持，这些流程和工具都将是空壳——实际上，OHDSI 社区的真正价值在于其成员。正如第 1 章所讨论的那样，OHDSI 社区的所有成员通过合作和开放科学，共同追求改善人类健康这一伟大愿景。

3.7 OHDSI 和 FAIR 指导原则

3.7.1 介绍

本章的最后一段使用 Wilkinson 等人出版的《FAIR 数据指导原则》，介绍了 OHDSI 社区和工具的当前现状。(2016 年)。

3.7.2 可寻性

从科学的角度来看，映射到 OMOP 并用于分析的任何医疗数据库都应该留存以备将来的参考和重现。而对 OMOP 数据库使用留存标识符的情况还不够普遍，部分是因为这些数据库通常部署在防火墙

后的内部网络上，而不一定连接到互联网上。但是，完全有可能将数据库摘要作为描述符记录发布，以作为 (例如引用目的的) 参考。如 EMIF 目录 17 即遵循了这种方法，它提供了数据库数据的收集目的、来源、词汇和术语、访问控制机制、许可、同意等方面的全面记录（Oliveira 等人，2019 年)。这种方法在 IMI EHDEN 项目中得到了进一步发展。

3.7.3 可及性

OMOP 映射数据的可及性是通过开放协议通常是 SQL 接口来实现的，该接口与 OMOP CDM 相结合，为访问 OMOP 数据提供了一种标准化、文档化的方法。然而，如上所述，出于安全原因，OMOP 数据源通常不能通过互联网直接获得。创建一个可供研究人员访问一个安全的全球医疗数据网络是一个积极的研究课题，也是像 IMI EHDEN 等项目的运营目标。然而，在多个 OMOP 数据库中的分析结果，如 LEGEND 和 http://howfund.org 等 OHDSI 计划所示，可以公开发表.

3.7.4 互操作性

互操作性可以说是 OMOP 数据模型和 OHDSI 工具的强项。为了在全球范围内建立一个强大的、可生成证据的医疗数据源网络，实现医疗数据源之间的互操作是关键，这是通过 OMOP 模型和标准词汇表实现的。然而，通过共享队列定义和统计方法，OHDSI 社区不仅仅实现了代码映射，还提供了一个平台来构建对医疗数据分析方法的深入理解。由于医院等医疗系统通常是 OMOP 数据的记录源，通过与 HL7 FHIR、HL7 CIMI 和 openEHR 等医疗保健操作互操作性标准保持一致，可以进一步增强 OHDSI 方法的互操作性。这同样适用于临床互操作性标准，如 CDISC 和生物医学本体。尤其是在肿瘤学等领域，这是一个重要的主题，OHDSI 社区的肿瘤学工作组和临床试验工作组是此类研讨论坛的典范。在引用其他数据，特别是本体术语方面，ATLAS 和 OHDSI-Athena 是重要的工具，因为它们允许在其他可用的医疗编码系统上下文中分析 OMOP 标准词汇表。

3.7.5 可复用性

围绕可复用性的 FAIR 原则集中在诸如数据许可、出处（阐明数据是如何存在）和与相关社区标准的链接等重要问题上。数据许可是一个复杂的话题，尤其是跨司法管辖区的数据许可，本书在此不进行广泛的介绍。但是，必须声明，如果您希望您的数据（例如分析结果）被其他人自由使用，通过数据许可证的形式来提供这些权限是一个很好的方法。令人遗憾的是，对于大多数可以在互联网上找到的数据来说，这还不是一种常见的做法， OHDSI 社区也不例外。关于 OMOP 数据库的数据源，可能的改进办法是使元数据以自动化方式被复用，包括如 CDM 版本、标准化词汇表发布、自定义代码列表等。OHDSI ETL 工具目前不会自动生成这些信息，但数据质量工作组和元数据工作组等组织正积极致力于这方面的工作。另一个重要方面是底层数据库本身的出处。了解医院或医生信息系统是否被替换或更改，以及历史上何时发生已知数据遗漏或其他数据问题，这些都是很重要的。探索如何在 OMOP CDM 中系统地附加入这些元数据，则是元数据工作组的工作领域。

- OHDSI社区可以被看作是一个开放科学社区，它积极地追求医学证据生成的互操作性和可重现性。
- OHDSI提倡从单一研究和单一评价结果的医学研究向大规模、系统性证据生成的范式转变。在这种模式下，基线发生率等事实是已知的。而新的证据侧重于利用来自真实世界的资源评价干预和治疗的效果。

参考文献

1. Allison, D. B., A. W. Brown, B. J. George, and K. A. Kaiser. 2016. "Reproducibility: A tragedy of errors." Nature 530 (7588): 27–29.
2. Chen, Xiaoli, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, et al. 2018. "Open Is Not Enough." Nature Physics 15 (2): 113–19. https://doi.org/10.1038/s41567-018-0342-2.
3. Garza, M., G. Del Fiol, J. Tenenbaum, A. Walden, and M. N. Zozus. 2016. "Evaluating common data models for use with a longitudinal community registry." J Biomed Inform 64 (December): 333–41.
4. Oliveira, José Luı´s, Alina Trifan, and Luı´s A. Bastião Silva. 2019. "EMIF Catalogue: A Collaborative Platform for Sharing and Reusing Biomedical Data." International Journal of Medical Informatics 126 (June): 35–45. https://doi.org/10.1016/j.ijmedinf.2019.02.006.
5. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. "Improving reproducibility by using high-throughput observational studies with empirical calibration." Philos Trans A Math Phys Eng Sci 376 (2128).
6. Wikipedia.2019a. "OpenscienceWikipedia,theFreeEncyclopedia." http://en.wikipedia.org/w/index.php?title=Open%20science&oldid=900178688.
7. Wikipedia.2019b. "Science2.0Wikipedia,theFreeEncyclopedia." http://en.wikipedia.org/w/index.php?title=Science%202.0&oldid=887565958.
8. Wikiquote.2019. "RonaldFisherWikiquote," \url{https://en.wikiquote.org/w/index.php?title=Ronald_Fisher&oldid=2638030}.
9. Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship." Sci Data 3 (March): 160018.

7. https://www.ohdsi.org/wp-content/uploads/2014/07/ARM-OHDSI_Duke.pdf
8. https://www.ehden.eu/webinars/
9. https://www.nature.com/collections/prbfkwmwvz
10. https://youtu.be/X5yuoJoL6xs
11.https://www.ema.europa.eu/en/events/common-data-model-europe-why-which-how
12. https://forums.ohdsi.org
13. https://www.ohdsi.org/web/wiki
14. https://www.ohdsi.org/web/wiki/doku.php?id=projects:overview
15. https://github.com/ohdsi
16. https://github.com/OHDSI/TheBookOfOhdsi
17. https://emif-catalogue.eu

第四章 通用数据模型

章节负责人: Clair Blacketer

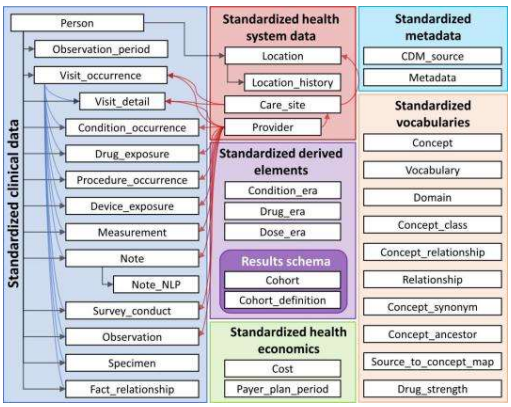
观察性数据可以反映患者接受诊疗期间的总体情况。全世界越来越多的患者在接受治疗期间产生的数据被收集和保存，从而形成了通常所说的“健康大数据”。 收集数据的目的包括三个方面：(1) 直接用于研究（通常以调查问卷或登记数据的形式)；(2) 支持诊疗活动的开展（通常称为 EHR-电子健康档案)；(3) 医疗费用管理（通常称为保险数据)。以上三类数据通常用于临床研究，其中后两类属于数据的二次利用，同时三类数据的格式和内容编码各不相同。

为什么我们需要一个用于观察性医疗数据的“通用数据模型”？

由于没有一个观察性数据库能够同等全面地记录患者在接受诊疗期间积累的所有临床事件。因此，必须从许多不同的数据源中提取研究结果，并进行比较和对比，以理解潜在记录偏倚的影响。同时，为了得出有统计学意义的结论，我们需要对大量患者进行观察研究，这也解释了同时评估和分析多个数据源的必要性。为此，不同的数据需遵循统一的通用数据标准进行治理。此外，患者数据需要严格的保护。传统上要提取数据用于分析，需基于严格的数据使用协议和复杂的访问控制。而使用通用数据标准可省略数据提取步骤，支持在本地环境中对数据进行标准化分析 - 让分析在数据端完成而不是先提取数据再去分析。

通用数据模型（Common Data Model，CDM）就是这样一种通用数据标准。CDM 与其标准化内容（参见第五章）相结合，可确保研究方法能被系统地应用，进而产生可比较和可重复的具有显著意义的结果。在本章中，我们将总体介绍数据模型、设计原理和使用约定，并对模型中的部分表进行讨论。CDM 中的所有表的概述如图 4.1 所示：

图 4.1： CDM 6.0 版本中所有表的概述（说明：此图并没有展示全部的表间关系）



4.1 设计原理

CDM 为支持典型的观察性研究目的而进行了优化：

识别某些进行了医疗干预（药物治疗、手术、医疗政策变化等）的患者群体和结局（疾病状况、手术、其他药物治疗等)；

根据不同的参数描述患者群体特征，如人口统计学信息、自然病史、诊疗服务、效用与成本、发病率、治疗方案和治疗路径等；

预测个体患者的医疗结局 - 参见第 13 章；

评价干预措施对人群的效果 - 参见第 12 章。

为了取得以上目标，CDM 的开发遵循以下原则：

适用性：CDM 旨在以最有利于分析的方式组织并提供数据，而不是为了满足医疗服务提供者或支付者的运营需求。

数据保护：所有有关患者身份和保密信息的数据都应受到限制，如姓名、准确的出生日期等。但如果研究明确需要更详细的信息，特殊情况下也可提供，例如对婴儿研究时需要准确出生日期。

域的设计：域（Domain）是在以人为中心的关系数据模型中建模的，其中每条记录至少要获取人的身份和日期。关系数据模型是将数据表示为由主键和外键链接的表集合的模型。

域的基本原理：如果域具有分析用例（如疾病状况）且具有其他方面不适用的特定属性，则在实体关系模型中会被标识并且被单独定义。所有其他数据可以作为观察值以“实体-属性-值”的结构保存在观察表中。

标准化术语表：CDM 采用标准化术语表来标准化这些记录中的内容，术语表涵盖了所有必要和适当的健康医疗相关概念标准。

复用现有术语表：如果可能，这些概念可以从国家、行业标准化组织、术语定义组织或计划中获得，如国家医学图书馆、退伍军人事务部、疾病控制和预防中心等。

维护源编码：即使所有的编码都被映射到了标准化术语表，CDM 模型仍会存储原始编码以确保不丢失任何信息。

技术中立：CDM 不需要特定技术。它可以在任何关系数据库中实现，如 Oracle、SQL Server 等，也可作为 SAS 分析数据集。

可扩展性：CDM 针对数据处理和计算分析进行了优化，以适应不同规模的数据源，包括含高达数亿人和数十亿临床观察案例的数据库。

向后兼容性：既往 CDM 所有的变更在 Github

存储库(https://github.com/OHDSI/CommonDataModel)中都有清晰的描述。旧版本的 CDM 可以从当前版本很轻易地创建，并且不会丢失先前存储的信息。

4.2 数据模型约定

CDM 中采用了大量隐含的和明确的约定。CDM 的研究者和开发者们需要理解这些约定。

4.2.1 模型的一般性约定

CDM 是一种“以人为中心”的模型，这意味着所有的临床事件表都与 PERSON 表相连接。将患者与事件日期或开始日期关联起来，我们就可以获得患者诊疗相关事件的纵视图。但标准化健康系统数据表并不适用于此规则，它直接链接到各个域的事件。

4.2.2 模式的一般性约定

在某些系统中，模式（Schemas）或数据库用户允许区分只读权限和读写权限。临床事件和术语表包含在 CDM 模式中，且对终端用户或分析工具来说是只读的。存储在“结果”模式中的表需要由基于 Web 的工具或终端用户操作。”结果”模式中的两个表为 COHORT(队列)和 COHORT_DEFINITION（队列定义）。这些数据表旨在存储用户自定义的目标患者群，详见第十章。这些表可以被写入，这意味着队列可以在运行时存储在 COHORT 表中。由于所有用户只有一个读写模式，因此如何组织和控制多用户的访问则取决于 CDM 实施中的具体设置。

4.2.3 数据表的一般性约定

CDM 是独立于平台的。通常使用 ANSI SQL 的数据类型进行定义（VARCHAR, INTEGER, FLOAT, DATE, DATETIME, CLOB）。只有 VARCHAR 会提供精确度，它反映了所需的最小字符串长度，但这可以在具体的 CDM 实例中扩展。CDM 并不对日期和时间的格式做要求。针对 CDM 的标准查询会因本地化实例及其日期/时间配置而异。

注：尽管 CDM 数据模型本身是独立于平台的，但在其之上开发的相当多的工具仍然需要遵照特定的规范，详见第八章。

4.2.4 域的一般性约定

不同性质的事件被组织到域（Domains）中。这些事件被存储在特定域的表和字段中，并由标准化术语表（参见 5.2.3 小节）中定义的特定领域的标准概念来表示。每一个标准化概念都被分配了唯一的域，它决定了此概念被记录在哪个表中。尽管域的分配在 OHDSI 社区中仍被不断探讨，但严格的“域-数据表-字段”的规则保证了任何代码或概念总是有一个明确的位置。例如，体征、症状和诊断概念属于疾病状况（Condition）域，相应的数据被存储在 CONDITION_OCCURRENCE 表的 CONDITION_CONCEPT_ID 字段。所谓的手术用药则通常以手术编码的形式记录在手术操作（Procedure）表中。在 CDM 中，这些记录存储于 DRUG_EXPOSURE（药物暴露表）中，因为他们映射到的标准概念归属于 Drug（药物）域。CDM 共有 30 个域，如表 4.1 所示：

表 4.1：每个域包含的标准概念数量（域名保留为英文名称）

概念数量	域ID	概念数量	域ID
1731378	Drug	183	Route
477597	Device	180	Currency
257000	Procedure	158	Payer
163807	Condition	123	Visit
145898	Observation	51	Cost
89645	Measurement	50	Race

33759	Spec Anatomic Site	13	Plan Stop Reason
17302	Meas Value	11	Plan
1799	Specimen	6	Episode
1215	Provider Specialty	6	Sponsor
1046	Unit	5	Meas Value Operator
944	Metadata	3	Spec Disease Status
538	Revenue Code	2	Gender
336	Type Concept	2	Ethnicity
194	Relationship	1	Observation Type

4.2.5 采用概念来表示内容

在CDM数据表中，每个记录的内容都是完全规范化的，并通过概念来表示。概念及对应的概念ID值共同存储在事件表中，概念ID是概念表的外键，其功能是用作通用参照表。所有CDM实例使用相同的CONCEPT（概念表）作为概念的参考，CONCEPT(概念表)与CDM同为互操作性的关键机制及OHDSI研究网络的基础。当一个标准概念不存在或无法识别时，则概念ID的值设置为0，表示不存在的概念、未知的概念或无法映射至概念的值。

概念表中记录了每一个概念的详细信息（包括名称、域、分类等）。标准化术语表包含了概念、概念关系、概念层级及其他概念有关的详细信息（参见第5章）.

4.2.6 字段命名的一般性约定

所有表中的变量名都遵循一个约定：

表 4.2: 字段名称约定

标记	描述
[Event]_ID	代表每一条记录的唯一标识符，作为外键和具体事件表建立关联。例如，PERSON_ID 唯一标识每一个患者或其他个体。VISIT_OCCURRENCE_ID 唯一标识每一次就诊。
[Event]_CONCEPT_ID	作为外键与 CONCEPT 参考表中的标准概念建立关联。[Event]_CONCEPT_ID 作为代表对应事件的主要标识，为后续的标准化分析奠

	定基础。例如，CONDITION_CONCEPT_ID = 31967 对应的参考值是 SNOMED CT 概念“Nausea”.
[Event]_SOURCE_CONCEPT_ID	作为外键与 CONCEPT 参考表中的概念建立关联。这里的概念指的是原始来源术语（示例如下），它可能恰巧是一个标准概念，有时候和[Event]_CONCEPT_ID 相同，它也可能是一个非标准概念。例如，CONDITION_SOURCE_CONCEPT_ID = 45431665 对应于 Read 术语表中的概念“Nausea”，进而其对应的标准概念则是 SNOMED CT 概念“31967”。可能存在某个事件歧义或语义不明而导致无法使用标准概念来表示的情况。此时可使用来源（非标准）概念支持进一步的分析和应用。我们不鼓励采用这种方式，因为来源（非标准）术语本身不能支持系统间的互操作。
[Event]_TYPE_CONCEPT_ID	作为外键与 CONCEPT 参考表中的记录建立关联。Type 型概念用来表示来源信息的出处，这些出处也以标准化概念的形式存储在标准术语表中。需要说明的一点是：尽管存储这个来源出处信息的字段并没有以“事件类型”或者“概念类型”来命名，但是这个字段存储了对应数据的采集方式。例如，DRUG_TYPE_CONCEPT_ID 可以区分一条药物记录是“非处方药”还是“处方药”。
[Event]_SOURCE_VALUE	来源编码或者来源字符串文本体现了事件在原始数据中的表达形式。由于在跨数据集情景下，来源形式的表达具有不一致性，因此，在用作标准化分析和应用时，我们不推荐使用来源术语来表示某事件。例如 CONDITION_SOURCE_VALUE 可能包含一个记录原始内容为“78702”，恰巧与忽略标记“.”之后的标准 ICD-9 编码 787.02 相同。

4.2.7 概念与源值的区别

在 CDM 中，相同的信息可能会存储在多个表的多个不同字段中，分别作为源值、来源概念或标准概念。

源值是指一个事件在其来源系统中的原始表达形式。一个事件的来源表达形式的标准化编码可能采用当前广泛应用的开放编码系统，例如 ICD-9-CM，NDC 或 Read，或是具有专属产权的编码系统例如 CPT4，GPI 或 MedDRA，亦或是来源系统中使用的受控词表（如 F 代表女性，M 代表男性），也可能是非标准的不受控的自由文本短句。源值存储在数据表对应的[Event]_SOURCE_VALUE 字段中。

概念是 CDM 特有的实体，是对临床事实的标准化表示。大多数概念都是基于现有的开放或具有专利的编码系统，而其他概念则是重新创建的（由“OMOP”开头的 CONCEPT_CODE）。 概念在所有域中都有唯一的 ID。

来源概念是指来源系统中使用的编码化概念。来源概念仅来自于现有的开放或专属编码系统，而不是 OMOP 生成的概念。来源概念存储在数据表的[Event]_SOURCE_CONCEPT_ID 字段中。

标准概念是用来定义临床实体意义的概念，在所有数据库中是唯一的，独立于来源编码系统。标准概念通常来自于现有的开放或具有专利的编码系统。在标准化术语表中，与标准概念具有相同含义的非标准概念被映射到对应的标准概念。标准概念（作为外键）在数据表的[Event]_CONCEPT_ID 字段中被引用。

提供源值只是为了方便查看和满足质量保证（QA）目的。源值包含的信息可能仅在特定数据来源的语义背景中有意义。如果来源系统中采用了某些标准编码系统，我们强烈推荐保留来源表达和来源概念，尽管这并不是必需的。而标准概念则是必需的。强制使用标准概念可确保所有 CDM 实例都可以使用相同的语言。例如，疾病“肺结核”的 ICD9CM 编码是 011（图 4.2）。

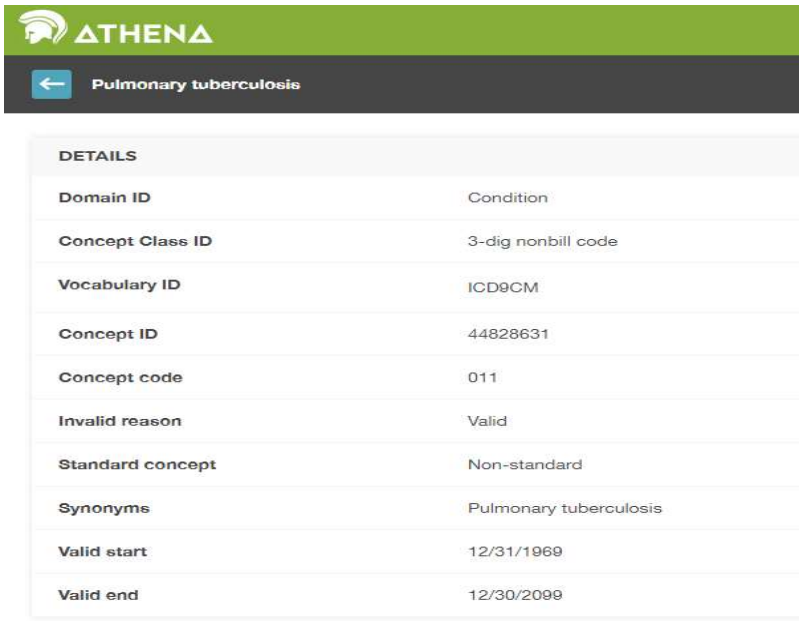


图 4.2: 肺结核（Pulmonary Tuberculosis）对应的 ICD9CM 编码

如果不考虑语义背景，编码 011 可能被认为是 UB04 术语表的“住院病人（包含 Medicare Part A)”，也可能被认为是 DRG 术语表的“无并发症或合并症的神经系统肿瘤”。这也就体现出来源概念 ID 及标准概念 ID 的重要价值了。ICD9CM 编码 011 对应的概念 ID 是 44828631，用不同的概念 ID 将 ICD9CM 与 UB04 和 DRG 的编码区分开来。如图 4.3 所示，ICD9CM 的肺结核来源概念通过“非标准到标准的映射(OMOP)”关系，映射到来自 SNOMED CT 术语表的标准概念 253954。Read、ICD10、

CIEL、MeSH 以及其他编码体系中也存在着相同的映射关系，因此，任一基于标准化 SNOMED CT 概念开展的 OHDSI 研究，也在一定程度上间接支持了所有的其他来源编码。

TERM CONNECTIONS (82)

RELATIONSHIP	RELATES TO	CONCEPT ID	VOCABULARY
ICD-9-CM to MedDRA (MSSO)	Pulmonary tuberculosis	36110777	MedDRA
Non-standard to Standard map (OMOP)	Pulmonary tuberculosis	253954	SNOMED
Subsumes	Other specified pulmonary tuberculosis	44830894	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination not done	44836741	ICD9CM
	Other specified pulmonary tuberculosis, bacteriological or histological examination unknown (at present)	44836742	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli found (in sputum) by microscopy	44821641	ICD9CM
	Other specified pulmonary tuberculosis, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture	44833188	ICD9CM

图 4.3: 肺结核（Pulmonary Tuberculosis）的 SNOMED 编码

表 4.7 展示了一个如何将非标准概念映射到标准概念的例子。

4.3 CDM 标准化表

CDM 包含 16 个临床事件表、10 个词汇表、2 个元数据表、4 个卫生系统数据表、2 个卫生经济学数据表、3 个标准化派生表和 2 个结果模式表。这些表已在 CDM Wiki.18 中详细论述。

为了说明在实践中如何使用这些表，本章的其余部分将用一个范例的数据全程阐明。

4.3.1 实践范例：子宫内膜异位症

子宫内膜异位症是一种痛苦的疾病，是指子宫内膜中的组织生长在子宫以外的部位。严重情况下可能引起不孕症、或肠道和膀胱的症状。以下各节将通过详细介绍一位该病患者的经历，来说明如何在通用数据模型中表示该疾病。

在这段痛苦旅程中的每一步，我都不得不向别人解释我是多么痛苦。Lauren 多年来一直有子宫内膜异位症的症状。然而，直到她的卵巢囊肿破裂后才得到诊断。您可以在 https://endometriosis-uk.org/laurens-story 上了解更多有关 Lauren 的信息。

4.3.2 患者（PERSON）表

我们对 Lauren 了解多少?

- 她是一位 36 岁的女性
- 她的生日是 1982 年 3 月 12 日
- 她是白种人
- 她是英国人

根据以上资料，她的患者（PERSON）表如下所示：

表 4.3: 患者表

列名	值	解释
PERSON_ID	1	PERSON_ID应该是一个整数，可以直接从源数据中获取，也可以在(CDM) 生产过程中产生。
GENDER_CONCEPT_ID	8532	女性的概念 ID 为 8532。
YEAR_OF_BIRTH	1982	
MONTH_OF_BIRTH	3	
DAY_OF_BIRTH	12	
BIRTH_DATETIME	1982-03-12 00:00:00	当时间未知时，使用午夜。
DEATH_DATETIME		
RACE_CONCEPT_ID	8527	白种人的概念 ID 是 8527，英国人的概念 ID 是 4093769，两者均正确，后者可以向上汇集到前者。请注意，在这里种族（Ethnicities）是作为人种（Races) 的一部分，而不是在ETHNICITY_CONCEPT_ID 中。

表 4.3: 患者表

列名	值	解释
ETHNICITY_CONCEPT_ ID	38003564	种族是美国区分西班牙裔和非西班牙裔的典型注释法，除美国外并不使用种族。此处病例为英国人，种族储存在 RACE_CONCEPT_ID 中。38003564 指“非西班牙裔”。
LOCATION_ID		她的地址未知。
PROVIDER_ID		她的家庭医生未知。
CARE_SITE		她的基层医疗单位未知。
PERSON_SOURCE_ VALUE	1	一般来说，这是她在源数据中的 ID，但它通常与 PERSON_ID 相同
GENDER_SOURCE_ VALUE	F	源数据中的性别值放在此处。
GENDER_SOURCE_CONCEPT_ID	0	如果源数据中的性别值是 OHDSI 支持的词汇代码，则这里是它的概念 ID。例如，如果她的性别在源数据中为“sex-F”，并且是 PCORNet 词汇，则此处为 Concept 44814665。
RACE_SOURCE_ VALUE	white	这里存放源数据中的人种的值。
RACE_SOURCE_ CONCEPT_ID	0	与 GENDER_CONCEPT_ID 的原理相同。

表 4.3: 患者表

列名	值	解释
ETHNICITY_SOURCE_VALUE	english	这里存放源数据中的种族的值。
ETHNICITY_SOURCE_CONCEPT_ID	0	与 GENDER_SOURCE_CONCEPT_ID 的原理相同。

4.3.3 观察期（OBSERVATION_PERIOD）表

当源系统中至少含有患者性别年龄、疾病状况、手术操作和药物的记录时，观察期（OBSERVATION_PERIOD）表旨在定义这些观察的时间跨度，其前提是有合理的灵敏度和特异性。对于保险数据，这通常是患者的参保期。电子健康档案则比较棘手，因为大多数医疗保健系统无法确定患者所去过的所有医疗机构或看过的所有医务人员。作为次佳解决方案，通常将系统中的第一条记录视为观察期的开始日期，而将最后记录视为结束日期。

如何定义 Lauren 的观察期（Observation Period）？

假设 Lauren 的信息如表 4.4 所示被记录在电子病历系统中，则她在观察期的医疗事件为：

表 4.4: Lauren 的医疗事件

就诊 ID	开始时间	结束时间	就诊类型
70	2010-01-06	2010-01-06	门诊
80	2011-01-06	2011-01-06	门诊
90	2012-01-06	2012-01-06	门诊
100	2013-01-07	2013-01-07	门诊
101	2013-01-14	2013-01-14	门诊
102	2013-01-17	2013-01-24	住院

根据医疗事件记录，她的观察期表应该如下所示：

表 4.5: 观察期表

列名	值	解释
OBSERVATION_ PERIOD_ID	1	通常是为每条记录创建的唯一标识符所自动生成的值。
PERSON_ID	1	PERSON_ID 是 Laura 在患者表记录的外键，将患者链接到观察期表。
OBSERVATION_PERIOD_ START_DATE	2010-01-06	最早的记录日期。
OBSERVATION_PERIOD_ END_DATE	2013-01-24	最晚的记录日期。
PERIOD_TYPE_ CONCEPT_ID	44814725	词汇表中“Obs Period Type”的最佳选择是 44814724，代表“就诊期”。

4.3.4 就诊（VISIT_OCCURRENCE）表

就诊（VISIT_OCCURRENCE）表包含有关患者在医疗卫生机构的活动信息。用 OHDSI 的方式来说，这些被称为就诊，属于独立事件。就诊具有广泛的层次结构，在最高处被分为 12 个类，分别表示不同的医疗服务方式。最常见就诊方式是住院、门诊、急诊和非医疗机构就诊。

如何用就诊来描述 Lauren 的经历?

作为示例，表 4.4 显示了我们如何将住院经历列入就诊表中。

表 4.6: 就诊表.

列名	值	解释
VISIT_OCCURRENCE_ ID	514	通常是为每条记录创建的唯一标识符所自动生成的值。
PERSON_ID	1	PERSON_ID 是 Laura 在患者表记录的外键，将患者链接到就诊表。
VISIT_CONCEPT_ID	9201	概念表的外键, 9201 代表住院。
VISIT_START_DATE	2013-01-17	就诊开始日期
VISIT_START_ DATETIME	2013-01-17 00:00:00	就诊开始日期和时间（时间未知，使用午夜）。
VISIT_END_DATE	2013-01-24	就诊结束日期。若就诊只有一天，则结束日期应与开始日期相同。
VISIT_END_DATETIME	2013-01-24 00:00:00	就诊结束日期和时间（时间未知，使用午夜）。
VISIT_TYPE_ CONCEPT_ID	32034	这里提供了有关就诊记录出处的信息，即它是否来自保险申报、医院账单或 EHR 记录等。在本示例中，使用概念 ID 32035（“EHR 的就诊记录”），因为对这些就诊的记录与

表 4.6: 就诊表.

列名	值	解释
		电子健康档案（EHR）相似。
PROVIDER_ID*	NULL	如果就诊记录含有医务人员的信息，则该 ID 记录在此处。这就是 PROVIDER 表中的 PROVIDER_ID。
CARE_SITE_ID	NULL	如果就诊记录含有医疗机构信息，则该医疗机构的 ID 放在这里。这就是 CARE_SITE 表中的 CARE_SITE_ID。
VISIT_SOURCE_ VALUE	inpatient	就诊的源数据值，Lauren 无该数据。
VISIT_SOURCE_ CONCEPT_ID	0	如果就诊的源数据值是 OHDSI 识别的词汇代码，则这里是它的概念 ID，Lauren 无该数据。
ADMITTED_FROM_ CONCEPT_ID	0	患者从何处入院的概念，此概念属于“就诊”域。如患者从家里住进了医院，则概念 ID 为 8536 “家”。
ADMITTED_FROM_ SOURCE_CONCEPT_ID	NULL	患者从何处入院的源数据值。按上述示例，此处取值为“家”。
DISCHARGE_TO_ CONCEPT_ID	0	患者出院后去处的概念，此概念属于“就诊”域。如患者出院后转到疗养院，则概念 ID 为 8615“疗养院”。

表 4.6: 就诊表.

列名	值	解释
DISCHARGE_TO_SOURCE_VALUE	0	患者出院后去处的源数据值。按上述示例，此处取值为“疗养院”。
PRECEDING_VISIT_OCCURRENCE_ID	NULL	表示本次就诊之前的就诊。与 ADMITTED_FROM_CONCEPT_ID 相比，它链接到实际的就诊记录而不是就诊概念。请注意，没有下一次就诊的链接时，就诊仅通过此列链接。

患者就诊期间，可能会与多个医务人员互动，住院期间尤为常见。这些互动可以记录在就诊详情（VISIT_DETAIL）表中。尽管本章未深入介绍，但您可以在 CDM wiki 中阅读有关就诊详情表的更多信息。

4.3.5 疾病状况（CONDITION_OCCURRENCE）表

疾病状况（CONDITION_OCCURRENCE）表记录了医务人员观察到的或患者描述的诊断、体征或症状。Lauren 的症状是什么？她回忆说：一直以来我都有痛经，大约三年前我感觉痛经变得越来越严重。我开始感觉到结肠附近有剧烈的戳刺性疼痛，尾骨和骨盆下方发软、肿胀，这导致了我每个月都有 1-2 天不能上班。止痛药有时能减轻疼痛，但通常没什么用处。

经期痛性痉挛，也称为痛经，SNOMED CT 编码为 266599000。表 4.7 显示了如何在疾病状况表中表示疾病状况：

表 4.7: 疾病状况表

列名	值	解释
CONDITION_OCCURRENCE_ID	964	通常是为每条记录创建的唯一标识符所自动生成的值。

PERSON_ID	1	PERSON_ID 是 Laura 在患者表记录的外键，将患者链接到疾病状况表。
CONDITION_ CONCEPT_ID	194696	SNOMEDCT 编码 266599000 的外键.
CONDITION_ START_ DATE	2010-01-06	症状开始日期。
CONDITION_ START_ DATETIME	2010-01-06 00:00:00	症状记录的日期和时间（时间未知，使用午夜）。
CONDITION_END_ DATE	NULL	症状结束日期，但很少记录。
CONDITION_END_ DATETIME	NULL	症状结束时间。
CONDITION_ TYPE_ CONCEPT_ID	32020	此列旨在提供有关记录出处的信息，即它来自保险申报、医院账单、EHR 记录等。在本示例中，使用概念 ID 32020 “EHR 就诊诊断”。 因为对这些就诊的记录与电子健康档案（EHR）相似。此列中的概念应该属于“疾病状况类型”词汇。
CONDITION_ STATUS_ CONCEPT_ID	0	症状记录时的状况。如入院诊断使用概念 ID 4203942。
STOP_REASON	NULL	源数据中所示的症状消失的原因。

PROVIDER_ID	NULL	如果症状记录中列出了做出诊断的医务人员，则该的 ID 记录在此处。这就是 PROVIDER 表中的 PROVIDER_ID，即当次就诊时的医务人员。
VISIT_OCCURRENCE_ ID	509	疾病状况得以诊断的当次就诊（VISIT_OCCURRENCE 表中 VISIT_OCCURRENCE_ID 的外键）。
CONDITION_SOURCE_VALUE	266599000	症状的源数据值。表示 Lauren 痛经的 SNOMED CT 编码记录在此处，表示该编码的概念放在 CONDITION_SOURCE_CONCEPT_ID，而它映射的标准概念则放在 CONDITION_CONCEPT_ID。
CONDITION_SOURCE_CONCEPT_ID	194696	如果使用 OHDSI 识别的词汇对症状的源数据值进行编码，则表示该值的概念 ID 记录在此处。在痛经的示例中，源值为 SNOMED CT 代码，因此表示该编码的概念为 194696。在这种情况下，它与 CONDITION_CONCEPT_ID 列中的值相同。
CONDITION_STATUS_SOURCE_VALUE	0	如果使用 OHDSI 支持的代码对源数据中的症状状况值进行编码，则该概念记录在此处。

4.3.6 用药记录（DRUG_EXPOSURE）表

用药记录（DRUG_EXPOSURE）表是有关计划或实际用药的记录，包括处方药、非处方药、疫苗和大分子生物疗法。用药记录可以从医嘱、处方、药房配药、住院用药以及其他患者主诉的相关的临床事件中得到。

如何描述 Lauren 的用药?

为了缓解痛经，在 2010 年 1 月 6 日的就诊时，医生给 Lauren 开了 60 片对乙酰氨基酚口服片剂

（又叫扑热息痛，在美国以 NDC 代码 69842087651 出售），每片 375 毫克，服用 30 天。在药物暴露用药记录表中显示如下：

表 4.8: 药物暴露用药记录表

列名	值	解释
DRUG_EXPOSURE_ID	1001	通常是为每条记录创建的唯一标识符所自动生成的值。
PERSON_ID	1	PERSON_ID 是 Laura 在患者表记录的外键，将患者链接到用药记录表。
DRUG_CONCEPT_ID	1127433	药品概念，对乙酰氨基酚的 NDC 编码映射到 RxNorm 的编码为 313782，而后者由概念 1127433 表示。
DRUG_EXPOSURE_START_DATE	2010-01-06	开始用药日期。
DRUG_EXPOSURE_START_DATETIME	2010-01-06 00:00:00	开始用药时间（时间未知，使用午夜）。
DRUG_EXPOSURE_END_DATE	2010-02-05	结束用药日期。根据不同的来源，它可能是一个已知的日期或一个推断的日期，表示患者用药的最后一天。由于我们知道 Lauren 有 30 天的药量，所以可以推断出这个日期。
DRUG_EXPOSURE_END_DATETIME	2010-02-05 00:00:00	结束用药的日期和时间。与 DRUG_EXPOSURE_END_DATE 的规则类似（时间未知，使用午夜）。
VERBATIM_END_DATE	NULL	源数据中明确记录的实际结束用药日期。推断的结束用药日期是基于患者使用了全部药量的假设。

表 4.8: 药物暴露用药记录表

列名	值	解释
DRUG_TYPE_CONCEPT_ID	38000177	记录来源的信息，包括来自保险申报、处方记录等。本示例使用概念 38000177“处方”。
STOP_REASON	NULL	停止用药的原因，包括完成疗程、更改药物、取消用药等。此信息很少被记录。
REFILLS	NULL	在许多国家，处方药系统中包含初始处方药后的自动补给数量。初始处方不计算在内，本值以 NULL 开头。对于 Lauren 的对乙酰氨基酚，她没有任何自动补给数量，因此该值为 NULL。
QUANTITY	60	原始处方或配药记录中的药量。
DAYS_SUPPLY	30	所开的药物供应天数。
SIG	NULL	原始处方或配药记录中的药品使用方法说明（“signetur”），在美国以药品处方的形式印在药品容器上。Signeturs 尚未在 CDM 中标准化，因而按原文输入。
ROUTE_CONCEPT_ID	4132161	该概念旨在表达患者的用药途径。Lauren 口服了对乙酰氨基酚，因此使用了概念 ID 4132161 “口服”。
LOT_NUMBER	NULL	药品厂家的产品批号。此信息很少被记录。
PROVIDER_ID	NULL	如果药品记录中列出了开处方的医务人员，则该 ID 记录在此处。这就是 PROVIDER 表中的 PROVIDER_ID。

表 4.8: 药物暴露用药记录表

列名	值	解释
VISIT_OCCURRENCE_ ID	509	开处方的那次就诊，即 VISIT_OCCURRENCE 表中 VISIT_OCCURRENCE_ID 的外键。
VISIT_DETAIL_ID	NULL	开处方的那次就诊，即 VISIT_DETAIL 表中 VISIT_DETAIL_ID 的外键。
DRUG_SOURCE_ VALUE	69842087651	药物在源数据中显示的源代码。在 Lauren 的例子中，此处为 NDC 代码。
DRUG_SOURCE_ CONCEPT_ID	750264	药物源数据值的概念。概念 750264 代表“对乙酰氨基酚 325 MG 口服片剂”的 NDC 代码。
ROUTE_SOURCE_ VALUE	NULL	源数据中关于给药途径的原文。

4.3.7 手术及操作（PROCEDURE_OCCURRENCE）表

手术及操作（PROCEDURE_OCCURRENCE）表包含医疗服务人员为诊断或治疗目的而对患者进行的手术及操作记录。手术及操作以不同的形式以及不同的标准化水平存在于各种数据源中。例如：

- 医疗保险记录中包含的手术或操作代码，是所申报的医疗服务项目的一部分。
- 电子病历中含有手术或操作的医嘱。

Lauren 有哪些手术及操作

根据 Lauren 的描述，我们知道她在 2013 年 1 月 14 日进行了左侧卵巢的超声检查，显示出 4x5cm 囊肿。在手术及操作表中显示如下：

表 4.9: 手术及操作表

列名	值	解释
PROCEDURE_ OCCURRENCE_ID	1277	通常是为每条记录创建的唯一标识符所自动生成的值。

表 4.9: 手术及操作表

列名	值	解释
PERSON_ID	1	PERSON_ID 是 Laura 在患者表记录的外键，将患者链接到手术或操作记录表。。
PROCEDURE_CONCEPT_ID	4127451	盆腔超声的 SNOMED 编码为 304435002，由概念 4127451 表示。
PROCEDURE_DATE	2013-01-14	操作日期。
PROCEDURE_DATETIME	2013-01-14 00:00:00	操作时间（时间未知，使用午夜）。
PROCEDURE_TYPE_CONCEPT_ID	38000275	有关操作记录出处的信息，即它是否来自保险申报、EHR 医嘱等。在本示例中，概念 ID 38000275 “EHR 医嘱录入”，用来表示操作来自电子病历记录。
MODIFIER_CONCEPT_ID	0	操作修饰词的概念 ID。如记录显示在身体双侧进行了 CPT4 操作，则使用概念 ID 42739579 “双侧操作”。
QUANTITY	0	手术或操作被开立医嘱或实际执行的数量。数量不存在、数量为 0 和 1 都代表同样意思。
PROVIDER_ID	NULL	如果手术或操作记录中列出了医务人员，则该 ID 记录在此处，即 PROVIDER 表中的 PROVIDER_ID。

表 4.9: 手术及操作表

列名	值	解释
VISIT_OCCURRENCE_ID	740	手术或操作的那次就诊（VISIT_OCCURRENCE 表中 VISIT_OCCURRENCE_ID 的外键）。
VISIT_DETAIL_ID	NULL	手术或操作的那次就诊细节。（即 VISIT_DETAIL 表中 VISIT_DETAIL_ID 的外键）。
PROCEDURE_SOURCE_VALUE	304435002	手术或操作在源数据中的编码。
PROCEDURE_SOURCE_CONCEPT_ID	4127451	手术或操作源数据值的概念。
MODIFIER_SOURCE_VALUE	NULL	源数据中修饰词的编码。

4.4 其他信息

本章仅涵盖了 CDM 的部分表作为示例。欢迎访问 Wiki 网站 19 获取更多信息。

4.5 总结

- CDM 旨在支持各种观察性研究。
- CDM 是以人为中心的数据模型。
- CDM 对数据结构进行标准化，并通过标准词汇表对数据内容的表现形式进行标准化。
- CDM 保留了源代码，以实现完全可溯源性。

4.6 练习题

预备知识

对于下面的第一套练习题，您需要查看前面讨论过的 CDM 表，并且可以通过 ATHENA20 或 ATLAS21 在词汇表中查找概念。

练习 4.1 John 是一名非裔美国人，出生于 1974 年 8 月 4 日。请在患者（PERSON）表中输入一个表达该信息的记录。

练习 4.2 John 在 2015 年 1 月 1 日参保。他的保险数据库中的数据在 2019 年 7 月 1 日被调取出来。请在观察（OBSERVATION_PERIOD）表中输入一个表达该信息的记录。

练习 4.3 2019 年 5 月 1 日，医生给 John 开了 30 天的布洛芬 200 毫克口服片（NDC 代码：76168009520）处方。请在药物暴露（DRUG_EXPOSURE）表中输入一个表达该信息的记录。

对于最后三个练习，除了需按第 8.4.5 节的说明安装 R、R-Studio 和 Java，您还需安装 SqlRender、DatabaseConnector 和 Eunomia 软件包，这些软件包可以使用以下方法安装：

```
install.packages(c("SqlRender",       "DatabaseConnector",       "devtools"))
devtools::install_github("ohdsi/Eunomia", ref = "v1.0.0")
```

Eunomia 软件包在 CDM 中提供了模拟的数据集，该数据集将在本地 R sessio 中运行。可以使用以下方法获取详细信息：

```
connectionDetails <- Eunomia::getEunomiaConnectionDetails()
```

CDM 数据库模式为“main”。

练习 4.4 使用 SQL 和 R，检索“胃肠道出血（Gastrointestinal hemorrhage）”状况的所有记录（概念 ID 为 192671）。

练习 4.5 使用 SQL 和 R，使用源代码检索“胃肠道出血（Gastrointestinal hemorrhage）”这一疾病状况的所有记录。该数据库使用 ICD-10，ICD-10 编码为“K92.2”。

练习 4.6 使用 SQL 和 R 检索 PERSON_ID 61 患者的观察期。

参考答案见附录 E.1。

Suggested answers can be found in Appendix E.1

18. https://github.com/OHDSI/CommonDataModel/wiki
19. https://github.com/OHDSI/CommonDataModel/wiki
20. http://athena.ohdsi.org/
21. http://atlas-demo.ohdsi.org

第五章 标准化术语集

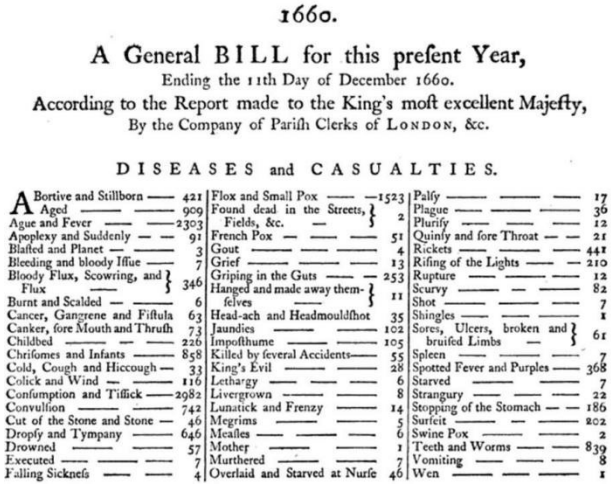
章节负责人：Christian Reich 和 Anna Ostropolets

OMOP 标准化术语集，通常被简称为“The Vocabulary”，是 OHDSI 研究网络的基础部分，也是通用数据模型（CDM）的组成部分。它通过规范数据内容来实现方法、定义和结果的标准化，为真正的远程（防火墙后）网络研究和分析奠定基础。通常情况下，观察性医疗数据会以统一编码的结构化数据或自由文本形式呈现，但数据的描述方式不尽相同。科研人员对数据进行整合、分析和挖掘的同时面临着描述方式不统一的难题。OHDSI 不仅要求数据格式上的统一，还要求数据内容上的统一。

在本章中，我们首先描述了标准化术语集的构建原则、组成部分、组成规则、常规用法和一些典型应用，这些都有利于理解和运用 CDM 通用数据模型。我们也会在其中明确术语集中继续完善的部分，希望得到相关科研机构的支持。

5.1 为什么需要术语集，为什么要进行标准化

医学术语集可以追溯到中世纪伦敦的死亡率统计表，用来管理鼠疫及其他疾病的爆发（见图 5.1）。



1660.

A General BILL for this preſent Year,

Ending the 11th Day of December 1660.

According to the Report made to the King's moſt excellent Majeſty,

By the Company of Pariſh Clerks of LONDON, &c.

DISEASES and CASUALTIES.

Abortive and Stillborn	421	Flox and Small Pox	1523	Palſy	17
Aged	909	Found dead in the Streets, Fields, &c.	2	Plague	36
Ague and Fever	2303	French Pox	51	Plurify	12
Apoplexy and Suddenly	91	Gout	4	Quinſy and ſore Throat	21
Blaſted and Planet	3	Grief	13	Rickets	441
Bleeding and bloody Iſſue	7	Griping in the Guts	253	Riſing of the Lights	210
Bloody Flux, Scowring, and Flux	346	Hanged and made away themſelves	11	Rupture	12
Burnt and Scalded	6	Head-ach and Headmouldſhot	35	Scurvy	82
Cancer, Gangrene and Fiſtula	63	Jaundies	102	Shot	7
Canker, ſore Mouth and Thruſh	73	Impoſthume	105	Shingles	1
Childbed	226	Killed by ſeveral Accidents	55	Sores, Ulcers, broken and bruiſed Limbs	61
Chriſomes and Infants	858	King's Evil	28	Spleen	7
Cold, Cough and Hiccough	33	Lethargy	6	Spotted Fever and Purples	368
Colick and Wind	116	Livergrown	8	Starved	7
Conſumption and Tiſſick	2982	Lunatick and Frenzy	14	Strangury	22
Convultion	742	Megrims	5	Stopping of the Stomach	186
Cut of the Stone and Stone	46	Meaſles	6	Surfeit	202
Dropſy and Tympany	646	Mother	1	Swine Pox	2
Drowned	57	Murthered	7	Teeth and Worms	839
Executed	7	Overlaid and Starved at Nurſe	46	Vomiting	8
Falling Sickneſs	4			Wen	1

图 5.1：1660 年伦敦死亡率统计表，使用当时已知的 62 种疾病分类系统显示了居民的死亡原因。

从那时起，面向医学词汇的分类方法已经在规模和复杂性上有了较大的提升，词汇内容涉及医疗保健的各个方面，例如诊疗操作、药物、医疗器械等。主要原则是一样的，包括一些医疗保健机构在对患者数据进行采集、分类以及分析时使用的受控词表、术语表、层级结构或本体。大部分术语集是由具有长期授权的公共机构和政府机构进行维护的。例如，世界卫生组织（WHO）编制的国际疾病分类法（ICD），最近已完成第 11 版（ICD11）的更新和发布。

各国政府还创建了具有符合本国国情特色的修订版本，如国际疾病分类法第十版临床修订版（美国）-ICD10CM、国际疾病分类法第十版临床修订版（德国）-ICD10GM 等。在对药品市场和药品销售过程实施管控的过程中，各国政府广泛使用医学术语集对药物销售、市场进行控制，并对国家认证的药物资源库进行维护。在私营部门，术语集可以作为一种商业产品，同时也支持内部使用，例如术语集可以用于电子健康档案（EHR）管理系统或医疗保险索赔报告中。

通常，每个国家、地区、医疗体系和机构都有各自的分类方法，并且他们使用的方法往往仅适用于他们自己。术语集数量繁多且形式各异，这就大大降低了系统间的互操作性。医疗信息标准化是实现全球范围内患者数据交换的关键，它使健康数据得以开放与共享，加快了系统化与标准化研究的进展，包括性能描述和质量评估。为了解决这一问题，跨国组织应运而生，并开始制定能够广泛应用的标准，如WHO、系统化临床医学术语表（SNOMED CT）和观测指标标识符逻辑命名与编码系统（LOINC）。在美国，卫生信息技术标准委员会（HITAC）建议国家卫生信息技术协调员（ONC）将 SNOMED CT、LOINC 和 RxNorm（药物词表）作为标准规范在一个通用平台上进行使用，以便实现全国医药卫生信息系统之间的数据交换。

OHDSI 开发了 OMOP-CDM，它是一个面向观察性医学研究的全球标准。OMOP 标准术语集是 CDM 模型的一部分，它有两个主要目的：

- 作为医学词汇的公共存储库，供各种机构使用
- 供研究者对数据进行标准化和映射

标准化术语集是向社会免费开放的，但必须作为必需参考表用于 OMOP CDM 实例。

5.1.1 构建标准化术语集

标准化术语集中的所有词汇都需要按照通用格式进行整合。因此，科研人员不必再对原始术语集的异构格式及使用规范进行处理和分析。所有词汇都通过 Pallas 系统完成定期更新和合并。系统由 OHDSI 术语集团队完成构建和后期维护，该团队是整个 OMOP CDM 工作组的一部分。如果您有任何建议或问题可以通过在 OHDSI Forums 或 CDM Github 页面留言，帮助我们不断改进和完善。

5.1.2 获取标准化术语集

标准化术语集可以直接通过 ATHENA 完成最新版本的下载并加载到本地数据库，不需要通过 Pallas 系统下载。同时，ATHENA 也支持术语集的分面搜索。

在下载标准化术语集的压缩文件时，请根据标准概念（见第 5.2.6 节）和常见用法来选择 OMOP CDM 模型中所有您所需的词汇，再添加源数据中使用的词汇。其中，专有词汇集没有勾选按钮。单击“获取许可”按钮，词汇集将会出现在您的列表中。随后，工作组将会与您联系，完成获取许可协议的认证；当您未获得许可协议时，工作组将为您提供相应的获取途径。

5.1.3 术语集来源：采用与构建

OHDSI 在术语集的选择上倾向于对已有术语集进行整合，而并非从头创建新的术语集，因为（i）许多术语集已经被应用于观察性数据的交换共享中，（ii）新词汇的构建和维护过程较为复杂，且需要投入大量人力。因此，在术语集构建过程中会有特定机构为工作组提供词汇表，所有概念全部遵循生成、弃用、整合和重用的生命周期（见第 5.2.10 节）。目前，OHDSI 只生成了内部管理术语集，如类型概念（包括疾病类型概念等）。RxNorm 扩展词集是一个例外，用于规范仅在美国境外使用的药物的命名（见第 5.6.9 节）。

5.2 概念

概念是用于描述 OMOP CDM 模型中临床事件的语义内容。这些概念将几乎所有表中的内容进行

了规范化的描述，是数据记录的基本组成模块。概念存储在 CONCEPT 表中（参见图 5.2）。

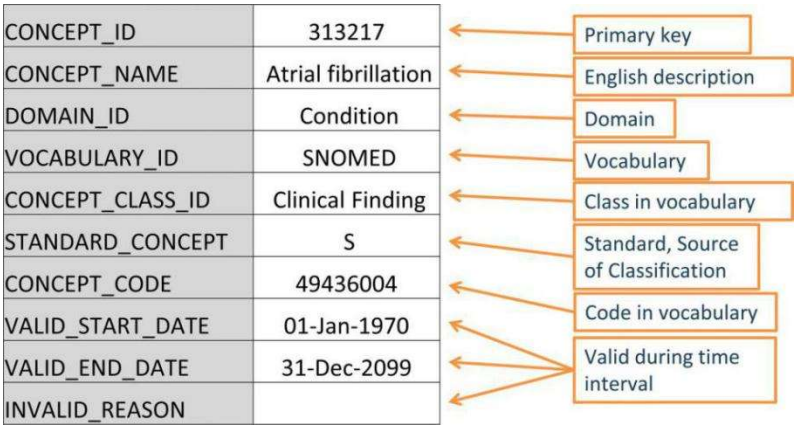


图 5.2： OMOP CDM 模型中概念的标准化示例。图例为 SNOMED CT 代码中“心房纤颤”的 CONCEPT 表。

这个系统是非常全面的，有足够多的概念可以覆盖患者就医过程中出现的任一事件（包括疾病症状、诊疗操作、药物治疗等）以及各类医疗卫生系统中的管理信息（包括就诊记录、护理记录等）。

5.2.1 概念唯一标识符

每个概念都被赋予一个唯一标识符（ID）作为主键。这个无意义整数 ID 用于标识 CDM 事件表中各类数据，并非来源词集的原始代码。

5.2.2 概念名称

每个概念都有一个名称。概念名称通常是从来源术语集中直接导入，用英文表示。如果来源术语集中有多个名称，则选择含义最丰富的名称作为概念名称；其余名称作为同义词存储在 CONCEPT_SYNONYM 表中的同一 CONCEPT_ID 下。非英文表示的名称也被存储在 CONCEPT_SYNONYM 表中，用相应的语言概念 ID 标注在 LANGUAGE_CONCEPT_ID 字段中。概念名称的最大长度为 255 字节，长度超过 255 字节的名称将被截取，而完整的概念名称将会作为其同义词进行存储，最大长度为 1000 字节。

5.2.3 类型域

每一个概念都会通过 DOMAIN_ID 字段赋予其一个类型域。与数值型的 CONCEPT_ID 不同，它可以包含数字也可以包含字母，且区分大小写，如“Condition”，“Drug”，“Procedure”，“Visit”，“Device”，“Specimen”等类型域的识别符号。当概念所属的类型域不明确或属于多个类型域时，可以出现组合域；通常标准概念（见第 5.2.6 节）是属于一个单一的类型域。类型域还表示临床事件/事件属性所属的 CDM 表或字段。类型域的分配是在词汇提取过程中 Pallas 系统进行的启发式推理，是 OMOP 的一个特定功能。来源术语集倾向于将不同的类型域按不同程度组合在一起（见图 5.3）。

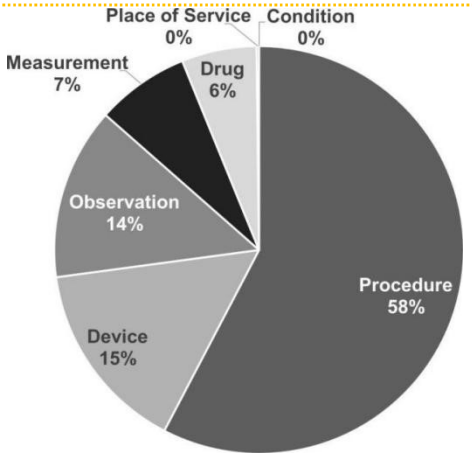


图 5.3：操作类术语集 CPT4 和 HCPCS 中类型域的分配。 按照常理，这些词汇包括其代码及概念应只属于同一个类型域，但实际上是归属于多个类型域。

启发式推理过程是由类型域的定义所决定的，即 CDM 模型中表和字段的定义（请参阅第 4 章）。目前，启发式推理过程尚待进一步完善，存在不明确的情况（请参见第 5.6 节“特殊情况说明”）。 如果您认为域的分配不正确，请在论坛或 CDM 问题帖子中留言，帮助我们改善分配流程。

5.2.4 术语集

每个术语集对应一个唯一标识符（包括字母和数字，且区分大小写），通常是在术语集的缩写基础上省略短破折号。例如，ICD-9-CM 的术语 ID 为“ICD9CM”。 OHDSI 目前涵盖了 111 种术语集，其中 78 种是通过外部来源采集，其余都是 OMOP 内部术语集。这些术语表通常每季度更新一次。词汇的来源和版本在 VOCABULARY 参考文件中进行了说明。

5.2.5 概念类型

一些术语集对其中的概念或代码进行了分类，并通过唯一标识符（区分字母大小写）来表示。例如，SNOMED CT 具有 33 个概念类型，是根据概念的垂直领域进行的划分，在 SNOMED CT 中称为“语义标签”，如临床发现、社会背景、身体结构等。其他术语集（例如 MedDRA 或 RxNorm 等）在其层级结构中对概念进行水平划分。没有概念类型的词集（例如 HCPCS）是将术语集 ID 当作概念类型 ID。

表 5.1：术语表的概念类型划分

概念类型划分原则	词表
水平划分	所有药物类词集，如 ATC, CDT, Episode, HCPCS, HemOnc, ICDs, MedDRA, OSM, Census
垂直划分	CIEL, HES Specialty, ICDO3, MeSH, NAACCR, NDFRT, OPCS4, PCORNET, Plan, PPI, Provider, SNOMED, SPL, UCUM

综合划分	CPT4, ISBT, LOINC
无划分	APC, all Type Concepts, Ethnicity, OXMIS, Race, Revenue Code, Sponsor, Supplier, UB04s, Visit

水平划分的概念类型位于层级结构的某个特定位置。例如，在药物术语集 RxNorm 中，概念类型“Ingredient”是位于层级结构的顶层。在垂直划分的概念类型中，每一个类型可以处于层次级别的任何位置。

5.2.6 标准概念

“标准概念”代表了每个临床事件的具体含义。例如，MESH 中的代码 D001281、CIEL 中的代码 148203、SNOMED CT 中的代码 49436004、ICD9CM 中的代码 427.31 和 Read 中的代码 G573000 都在疾病类型域中定义了“心房纤颤”这一概念，但只有 SNOMED CT 中的概念被指定为标准概念，用来指代该疾病。其他代码则作为非标准概念或来源概念，并可以映射到标准概念上。标准概念在 STANDARD_CONCEPT 字段中以“S”表示。只有标准概念才可以用于在 CDM 字段中记录数据，以“_CONCEPT_ID”结尾。

5.2.7 非标准概念

非标准概念不用于表示临床事件，但是它们仍然是标准化术语集的一部分，通常在来源数据中可以找到。因此，它们也被称为“来源概念”。从非标准概念到标准概念的转换是一个 “映射”的过程（请参见第 5.3.1 节）。 非标准概念在 STANDARD_CONCEPT 字段中没有值（NULL）。

5.2.8 分类概念

分类概念不是标准概念，因此不能用于表示数据。但是分类概念和标准概念一起参与层级结构的表示，能够支持层级结构的查询。例如，查询 MedDRA 中代码 10037908 的所有子结点（对于未获得 MedDRA 许可证的用户不可见，有关访问限制的信息请参阅第 5.1.2 节）将对心房纤颤的 SNOMED CT 标准概念进行检索（使用 CONCEPT_ANCESTOR 表进行分层查询，请参阅第 5.4 节）-见图 5.4。

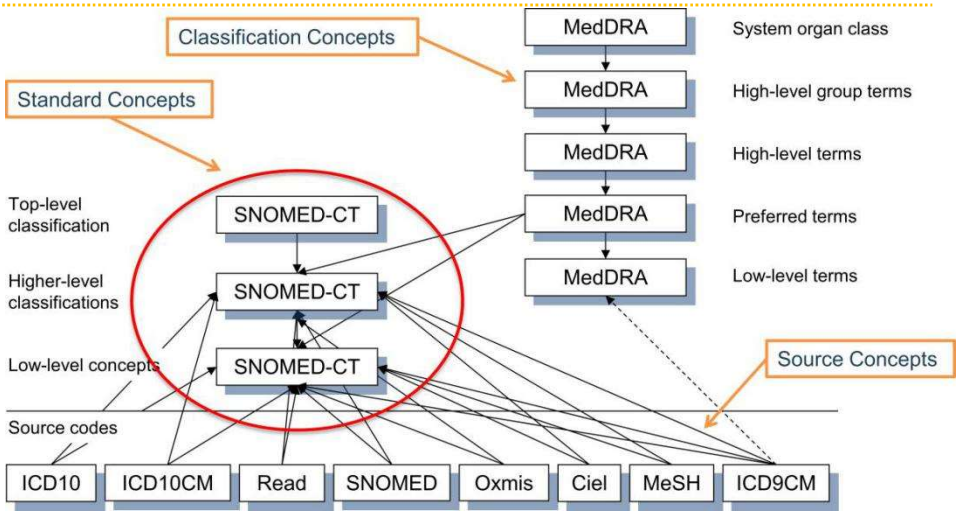


图 5.4：疾病域中的标准概念、非标准概念、分类概念及层级关系。

SNOMED CT 被指定为标准概念（其中包含一些与 ICDO3 相关的肿瘤学概念），MedDRA 作为分类概念参与层级结构的表示，其余术语集包含了非标准概念或源概念，不参与层级结构的表示。

标准概念、非标准概念和分类概念在术语集层面的选择通常是针对每个类型域单独进行的。它是基于术语集中概念的质量、层级结构的设计以及该词集的应用范围。另外，并非所有术语集中的概念都会被当作标准概念。针对不同的类型域，每个术语表所属的概念类型是不同的，且术语表中的概念都必须处于有效状态（请参阅第 5.2.10 节）；如果不同术语集中的多个概念对应到同一个含义，会对概念进行优先排序。也就是说，没有一个固定“标准词汇集”。标准概念的划分，请参见表 5.2。

表 5.2： 医学术语集被用作标准/非标准/分类概念的情况列表

域	标准概念	来源概念	分类概念
病症	SNOMED CT, ICDO3	SNOMED,CT Veterinary	MedDRA
诊疗操作	SNOMED CT, CPT4,	SNOMED CT, CPT4,	无
检测	SNOMED CT, LOINC	SNOMED CT Veterinary,	无
药物	RxNorm, RxNorm Extension, CVX	RxNorm, RxNorm Extension, CVX	ATC
医疗器械	SNOMED CT	其他，目前还未进行标准化	无
观测	SNOMED CT	其他	无

就诊访问	CMS Place of Service, ABMT, NUCC	CMS Place of Service, ABMT, NUCC	无

5.2.9 概念代码

概念代码是来源术语集中使用的标识符。例如，ICD9CM 或 NDC 的代码会存储在这个字段中。因为同一个代码在不同的术语集中可能代表不同的含义（请参见表 5.3），因此 OMOP 表在使用概念 ID 时是将其作为 CONCEPT 表的外键。

表 5.3：具有相同概念代码 1001 的概念，但其术语、类型域和概念类型却不一样。

概念 ID	概念代码	概念名称	类型域 ID	术语 ID	概念类型
35803438	1001	Granulocytecolony-stimulating factors	Drug	HemOnc	Component Class
35942070	1001	AJCC TNM ClinT	Measurement	NAACCR	NAACCR
1036059	1001	Antipyrine	Drug	RxNorm	Ingredient
38003544	1001	ResidentialTreatment-Psychiatric	Revenue Code	Revenue Code	Revenue Code
43228317	1001	Aceprometazine maleate	Drug	BDPM	Ingredient
45417187	1001	Brompheniramine Maleate,1mg/mL injectable solution	Drug	Multum	Multum
45912144	1001	Serum	Specimen	CIEL	Specimen

5.2.10 生命周期

术语集是较少见的具有一套固定代码集合的永久性语料库。使用及维护过程中，可以对概念和代码进行增加和删减。由于 OMOP CDM 模型可用于患者纵向长期数据的规范化，因此模型必须包含曾经使用但可能不再有效的概念，也必须支持概念的新增并将其置于上下文中。CONCEPT 表中有三个字段可以描述概念的生命周期：VALID_START_DATE，VALID_END_DATE 和 INVALID_REASON。 字段取值根据概念所处的生命周期而异：

- 有效或新概念
  - 说明：使用中的概念。
  - VALID_START_DATE：概念实例化的日期；如果未知，则使用概念编入词汇集的时间；如果仍未知，则默认为 1970-1-1。
  - VALID_END_DATE：初始值设置为 2099-12-31，表示“可能在将来失效，但现在有效”。

– INVALID_REASON: NULL

- 已弃用的概念，且没有后继概念

– 说明：概念已无效，不能用作标准概念（请参阅第 5.2.6 节）。

– VALID_START_DATE：概念实例化的日期；如果未知，则使用概念编入词汇集的时间；如果仍未知，则默认为 1970-1-1。

– VALID_END_DATE：概念被弃用的日期；如果未知，则使用概念丢失或无效后词集完成更新的日期。

– INVALID_REASON: “D”

- 完成升级的后继概念

– 说明：概念本身已失效，但已定义了后继概念。删除重复性数据是其中的一种典型情况。

– VALID_START_DATE：概念实例化的日期；如果未知，则使用概念编入词汇集的时间；如果仍未知，则默认为 1970-1-1。

– VALID_END_DATE：概念完成升级的日期；如果未知，则使用概念完成升级后词集完成更新的日期。

– INVALID_REASON: “U”

- 复用代码的新概念

– 说明：将词汇集中弃用概念的代码复用于一个新概念。

– VALID_START_DATE：概念实例化的日期；如果未知，则使用概念编入词汇集的时间；如果仍未知，则默认为 1970-1-1。

– VALID_END_DATE：概念被弃用的日期；如果未知，则使用概念丢失或无效后词集更新的日期。

– INVALID_REASON: “R”

一般来说，概念代码是不可重复使用的。但也有一些术语集未遵循这一规则，例如 HCPCS、NDC 和 DRG。这些术语集中会存在同一个概念代码对应多个概念的情况。其 CONCEPT_ID 是保持了独特性的。复用概念代码在 INVALID_REASON 字段中用“R”标记，并使用有效的 VALID_START_DATE 到 VALID_END_DATE 区间来区分使用同一个概念代码的多个概念。

5.3 关联性

无论两个概念是否隶属于同一个类型域或者术语集，它们之间都有一个固定的关系。这种固定关联的属性是通过 CONCEPT_RELATIONSHIP 表中 RELATIONSHIP_ID 字段来表达的。该字段 ID 可以用字母或数字来表述，长度很短，需要区分大小写，并且每一个编号都是独特的。每一对关系都是对称的，即与自身的反向关系完全对等：字段 CONCEPT_ID_1 与字段 CONCEPT_ID_2 的内容互换，并且 RELATIONSHIP_ID 是颠倒的。例如，关系“映射到”的反向关系为“从...映射”。

CONCEPT_RELATIONSHIP 表里也有关于生命周期的字段，如 RELATIONSHIP_START_DATE，RELATIONSHIP_END_DATE 以及 INVALID_REASON。但是只有有效记录，即当字段 INVALID_REASON 不为 NULL 的时候，此条记录才会出现在 ATHENA 中。非有效状态的记录会保存在 Pallas 系统中，仅用于内部存储和处理。RELATIONSHIP 表格记载了所有关联性 ID 及他们所对应的反向关系，用以参考。

5.3.1 映射关系

映射关系由两对关联性 ID 构成（见表 5.4），将非标准化的概念转化为标准概念。

表 5.4：映射关系的类型

关联性 ID 组合	目的
“映射到”和“从...映射”	转化为标准概念。如果一个概念本身已经是标准的，那么就映射为它自己；否则要映射为标准概念。大多数非标准化概念和所有的标准概念都有这样的属性。映射前的字段存储在*_SOURCE_CONCEPT_ID，映射后的字段存储在*_CONCEPT_ID 中。用来分类的属性字段无映射关系。
“映射到数值”和“从...映射而来的数值”	将原有的概念用一个数值来表示。这个数值将被储存于 MEASUREMENT 和 OBSERVATION 表格中的 VALUE_AS_CONCEPT_ID 字段。

这些映射关系在相等的概念之间搭建起了桥梁，互通有无，使得临床医疗事件在 OMOP CDM 中的表达更加统一。这是标准化术语集的主要成就之一。

“相等概念”指含义相同的概念，且更重要的是，它们的子级层级必须覆盖同样的语义空间。如果某一个相等概念是缺失的，并且概念是非标准化的，那么它会被映射到一个含义更加宽泛的概念当中，这个过程又称为“向上映射”。举个例子，ICD10CM W61.51“被鹅咬了”在 SNOMED CT 术语集中没有对应的相等概念（SNOMED CT 常作为统一标准下的语义概念），取而代之，这个概念被映射为 SNOMED CT 21776004 “被鸟啄了”，而后者忽略了鸟是鹅的这层含义。向上映射仅适用于与规范研究案例无关的信息缺失。

在某些映射关系中，一个来源概念可以对应一个以上的标准概念。比如 ICD9CM 070.43“戊型肝炎并发肝昏迷”被映射到 SNOMED CT 235867002“急性戊型肝炎”以及 SNOMED CT 72836002“肝昏迷”。这是因为来源概念是由两个症状共同组成的，即肝炎和昏迷。因为 SNOMED CT 不包含这样的组合，所以 ICD9CM 映射后出现了两条记录，每条记录都有相对应的标准概念。

“映射到数值”这种关系的目的是根据实体-属性-数值（EAV）模型把一个数值分离出来后加入到 OMOP CDM 表格中。这种应用往往出现在以下场景：

- 包含一个检测值和一个结果值的度量
- 个人或家庭疾病史
- 对于某种物质的过敏
- 对于疫苗注射的需求

在这些情况中，来源概念均由属性（如：检测或历史）和数值（如：检测值或疾病）组合而成。关系“映射到”可以为来源概念找到与其对应的属性，而“映射到数值”可以找到相应的数值。具体例子

参见图 5.5。

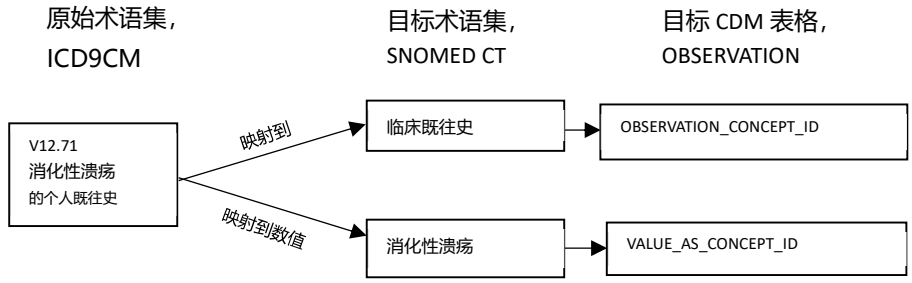


图 5.5：来源概念与标准概念之间的一对多关联映射。

一个先组配型概念被划分为两个不同的概念，其中一个是属性（如临床既往史），另一个是数值（如消化性溃疡）。关系“映射到”对应的是测量或观察到的属性，而“映射到数值”没有任何类型域方面的限制。

5.3.2 层级关系

象征层级结构的关联关系是通过“属于”—“将...归入”关系组合来定义的。层级关系的定义是：子概念具有父概念的所有属性，且子概念新增了一个或多个属性或对某个属性进行了更为精准的定义。例如，SNOMED CT 49436004 “心房纤颤”与 SNOMED CT 17366009 “房性心律失常”之间的关联关系是“属于”。除了心率失常的类型以外，这两个概念的属性几乎完全一致。其中一个概念将心律失常定义为“纤颤”，而另一个概念没有明确。所有语义概念均可以有一个以上的父概念或子概念。在上面的例子中，SNOMED CT 49436004 “心房纤颤” 也同样“属于”另一个概念 SNOMED CT 40593004 “纤颤”。

5.3.3 不同术语集的概念之间的关联关系

通常情况下，不同术语集的概念之间的关系是“词汇 A — 词汇 B 相等”。这种关联关系一部分来自于词汇的原始语库，另一部分是由 OHDSI 术语集团队搭建的。它们可以提供近似的映射，但准确度经常低于经过周密映射的关联关系。高质量的相等关系（如“Source-RxNorm 相等”）往往会在关系“映射到”中重复出现。

5.3.4 同一术语集的概念之间的关联关系

术语集内部的关联关系普遍来自词汇的原始提供方。完整描述详见 OHDSI Wiki 中独立术语集的词汇文件。

这类关联关系中，大部分都用来定义临床事件之间的关系，并且便于信息检索。例如，从“查找...的病灶”关系中可以搜索到尿道疾病（见表格 5.5）：

表 5.5：概念“尿道”的“查找...的病灶”关系，其包含的所有病症都来自解剖学结构。

CONCEPT_ID_1	CONCEPT_ID_2
4000504 “尿道部位”	36713433 “尿道部分重复”

4000504 “尿道部位”	433583 “尿道上裂”
4000504 “尿道部位”	443533 “尿道上裂，男性”
4000504 “尿道部位”	4005956 “尿道上裂，女性”

这些关联关系的质量和完整性取决于原始术语集的质量。这种情况下，一般会使用那些可以获取标准化概念的术语集，比如 SNOMED CT。由于进行了更好的校验，它们拥有质量更高的术语集内部的关联关系。

5.4 层级结构

在同一个类型域里，标准概念和分类概念均统一组织编排于一个层级结构中，且储存于 CONCEPT_ANCESTOR 表格。这样便于查询和检索获取不同概念以及它们所有的后裔概念。这些后裔概念的属性与它们的先祖一致，但会出现新增或定义更加精准的属性。

CONCEPT_ANCESTOR 表格是从 CONCEPT_RELATIONSHIP 表格中自动生成的，后者将所有潜在的概念都通过层级关系互相连接。这些层级关系包含了“属于”—“将...归入”这样的组合（见图 5.6），也包括其它一些将不同术语集之间的层级结构关联起来的关系。如果想了解某个关联关系是否参与到层级结构中，可以查看该关系 ID 在 RELATIONSHIP 参考表格中的重要标示 DEFINES_ANCESTRY。

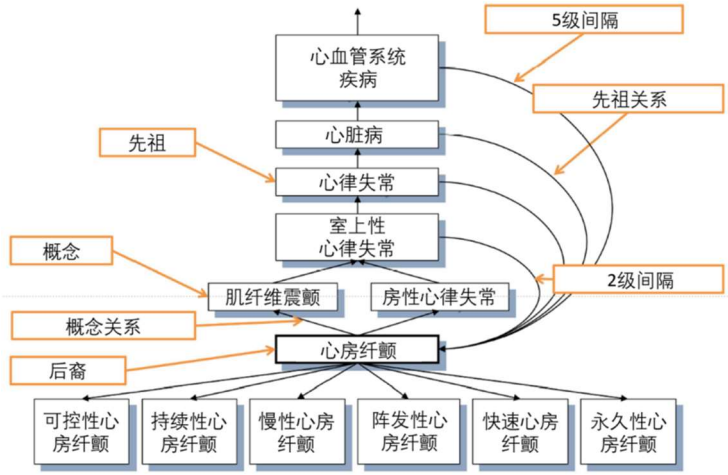


图 5.6：疾病“心房纤颤”的层级结构。

第一级先祖通过“属于”和“将...归入”的关系来定义，而所有层级更高的关联关系都可以从 CONCEPT_ANCESTOR 表格中推测出来，并储存于该表格中。除了已有的其它后裔，任何概念同时也是其自己的后裔，且两者相间隔的级数为 0。

MIN_LEVELS_OF_SEPERATIONMAX_LEVELS_OF_SEPARATION 分别定义了最短和最长的先祖级数，即先祖和后裔概念之间所间隔的层级数量。并非所有层级关系在分离程度的计算过程有同等的贡献值。RELATIONSHIP 参考表格中的标示 IS_HIERARCHICAL 决定了所有关联关系 ID 的层级数量。

截至目前，一个高质量的完整层级结构仅仅出现在两个类型域中：药物和疾病。而该架构仅覆盖了其他类型域的一小部分，如诊疗操作、检测数值和医学观测值。先祖概念在药物域起着举足轻重的作

用，主要原因是它可以将药物成分和药物类型从原产国、商品名等其它属性中分离出来，易于根据前两者来查找所有药物。

5.5 内部参考表

部分字段由它们自己的术语集控制，如 DOMAIN_ID, VOCABULARY_ID, CONCEPT_CLASS_ID（全部储存于 CONCEPT 表中）以及 CONCEPT_RELATIONSHIP_ID（储存于 CONCEPT_RELATIONSHIP)。它们定义可以在四个参考表格中查看，包括 DOMAIN, VOCABULARY, CONCEPT_CLASS 和 RELATIONSHIP。在这些参考表格中，字段*_ID 是主键；字段*_NAME 涵盖更细节的描述；而字段*_CONCEPT_ID 引用了 CONCEPT 表格，后者为每一个参考表格中的记录提供了对应的概念。这些重复的记录旨在驱动信息模型的自动化运行。另，表格 VOCABULARY 包含了字段 VOCABULARY_REFERENCE 和 VOCABULARY_VERSION，两者分别指代原始术语集的出处和版本。表格 RELATONSHIP 还纳入了其它字段，包括 DEFINES_ANCESTRY, IS_HIERARCHICAL 以及 REVERSE_RELATIONSHIP_ID。最后的这个字段 REVERSE_RELATIONSHIP_ID 定义了每一对反向关联关系 ID。

5.6 特殊情形

5.6.1 性别

在 OMOP CDM 和标准化术语集中，性别指出生时的生理性别。而真正亟待解决的问题是如何定义那些非传统意义上的性别。这些应用场景必须由 OBSERVATION 表中的记录来实现。若有相关信息，所有用户自定义的性别均储存于该表格中。

5.6.2 种族和民族

种族和民族的定义遵循了美国政府的相关规定。美国拉丁裔人群和非拉丁裔人群拥有同样的种族，但他们的民族却不一致。种族可以分为 5 个常见的类别，而民族则是它们在层级结构中的后裔。该结构不包含混血种族。

5.6.3 诊断编码体系和 OMOP 疾病

人们常使用的编码体系，如 ICD-9 或 ICD-10，是根据合理的病情检查来确诊的，因此该诊断编码的精准度会出现差异。这个语义空间和疾病症状域并不完全一致，但会有部分重叠。举个例子，疾病症状可以包含在确诊之前就记录下的体征和症状，而 ICD 编码还包含了隶属于其它类型域的概念（如：诊疗操作)。

5.6.4 操作编码系统

同理，类似于 HCPCS 和 CPT4 的编码体系通常被认为是医学操作的目录。而实际上，它们其实为医疗服务的支付方式提供了合理的解释。诊疗操作域涵盖了大部分的医疗服务，但也有许多概念是诊疗操作没有涉及到的。

5.6.5 医疗器械

没有一个统一的编码体系可以用来规范化关于医疗器械的概念。在许多原始数据中，医疗器械还未被纳入外部编码体系。出于这个原因，目前暂时没有相关的层级结构系统。

5.6.6 就诊与诊疗服务

就诊的概念定义了与医疗服务中病人就医行为相关的事件。在许多源系统中，它们被称作就诊地点，指代了一些机构或实体建筑，比如医院。而在其它系统中，它们被称为诊疗服务。在不同国家之间，这些名称有所差异，并且它们是很难被准确定义的。诊疗场所往往会根据数次就诊中的一次来明确其取值（如：医院 XYZ），但在研究中不能简单将医院 XYZ 定义为唯一诊疗场所，比如在 XYZ 医院接受诊疗服务病人也可能发生其他诊疗场所的就诊。

5.6.7 提供方与专科

任何一个能够提供服务的人士都属于提供方的范畴。他们包括医务工作者，如医生和护士，但也包括了非医疗服务的提供者，如验光师和鞋匠。专科是隶属于提供方“医生”的后裔概念。虽然人们经常依据主体员工的专业类型域来命名其所在的诊疗场所（如：“外科”），但是诊疗场所并不具备专科这一概念。

5.6.8 具有特殊要求的治疗领域

标准化术语集完整地覆盖了医疗健康的各个方面。但是有的治疗领域具备特殊的要求，因此需要专门针对它们设计独特的术语集。这些领域包括但不局限于肿瘤学、放疗和基因组学。相应的扩展词表由 OHDSI 特别工作小组开发。因此，OMOP 标准化术语集将不同来源和用途的概念汇集到统一的具有域特定性的层级结构，进而成功建立了一个集成系统。

5.6.9 药物域的标准化概念

大多数药物域的概念都是从美国国立医学图书馆的一个公开术语集 RxNorm 中获取来的。然而很多美国境外的药品不一定能够被该术语集覆盖，这取决于药品的成分、剂型和规格的组合是否已经在美国上市。OHDSI 术语集团队把尚未在美国上市的药品添置于 RxNorm 扩展词集中，这是 OHDSI 唯一一个自主创建的大型类型域术语集。

5.6.10 空值字段 NULL

许多术语集的编码都会出现缺失的信息。例如，在以下五个性别概念中，8507 “男性”，8532 “女性”，8570 “尚未明确”，8551 “未知”，以及 8521 “其他”，仅有前两者是标准化的，剩下三个都是没有映射对象的原始概念。标准化术语集并没有明确某一信息缺失的原因；可能是由于病人主动撤回自己的隐私信息，或缺乏原始数值。

或这些数值没有被精确或规范化，也可能因为：

CONCEPT_RELATIONSHIP 表格中缺少了对应的映射记录。以上任何一个原因导致的缺失概念都会被默认映射为标准化概念中的 ID=0。

5.7 总结

- 所有事件和管理规则均体现在 OMOP 标准化术语集中，它们的表现形式包括概念、概念间的关系和概念的先祖层级结构。
- 这些表达方式大部分都来自于现存的编码体系或术语集，但也有部分是由 OHDSI 术语集团队二次开发的。
- 所有概念都属于某一特定的类型域，该域限定了相关概念所表达的事实在 CDM 中储存的位置。
- 如果来自于不同术语集的概念具有相同的含义，那么其中一个概念会作为标准化概念被映射，而其余的作为来源概念。
- 映射这个步骤是通过概念关系“映射到”和“映射到数值”来完成的。
- 另有一种类型的概念叫做分类概念。它们是非标准化的，并且不同于来源概念，它们是层级结构的一部分。
- 所有概念都有生命周期，是与时俱进的。
- 同一个类型域中的概念被统一编排进层级结构中。不同类型域之间，层级结构的质量会参差不齐。完善层级结构系统将是一个持续性的工作 。
- 如果你发现了错误或者不准确的表达，我们鼓励你和 OHDSI 社区进行联系。

5.8 练习

答题要求

在回答以下题目时，你需要查看标准化术语集中的概念。你可以使用 ATHENA 或 ATLAS 来完成。

练习 5.1. “消化道出血”的标准化概念 ID 是什么?

练习 5.2. 哪一个 ICD-10CM 的编码映射到了“消化道出血”的标准化概念？哪一个 ICD-9CM 的编码映射到了这个标准化概念?

练习 5.3. 有哪些 MedDRA 中的优选术语与“消化道出血”的标准化概念是相等的?

参考答案详见附录 E.2。

22. https://github.com/OHDSI/Vocabulary-v5.0
23. https://forums.ohdsi.org
24. https://github.com/OHDSI/CommonDataModel/issues
25. http://athena.ohdsi.org
26. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:mapping
27. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary
28. http://athena.ohdsi.org/

第 6 章 ETL 技术

章节负责人：Clair Blacketer & Erica Voss

6.1 简介

为了从本地/原始数据转换到 OMOP 通用数据模型（CDM），我们需要创建一个提取-转换-加载（ETL） 过程。此过程应将数据重组成 CDM 结构，并向标准术语集 （Standardized Vocabularies）添加映射。该过程通常以一组自动化脚本（例如 SQL 脚本）的形式实现。ETL 过程的可重复性至关重要，以便每当源数据被更新时可以重新运行。

创建一个 ETL 任务通常非常艰巨。多年以来，我们开发了包括四个主要步骤的最佳实践：

1.数据专家和 CDM 专家共同设计 ETL 任务。

2.具有医学知识的人创建代码映射。

3.技术人员实施 ETL。

4.所有人都参与质量控制。

在本章中，我们将详细讨论每个步骤。OHDSI 社区已经开发了一些工具来支持其中一些步骤，本章也会讨论这些工具。在本章结束时，我们将讨论 CDM 和 ETL 的维护。

6.2 步骤 1：设计 ETL

首先要将 ETL 的设计与 ETL 的实现明确区分开。设计 ETL 需要对源数据和 CDM 都有广泛的了解，而实现 ETL 通常主要依赖于如何提高 ETL 的计算效率。如果我们试图同时做这两件事，很可能会陷入细节上的泥潭，所以应该专注于整体情况。我们已经开发了两个紧密集成的工具来支持 ETL 设计过程：White Rabbit 和 Rabbit-in-a-Hat。

6.2.1 White Rabbit

要在数据库上启动 ETL 过程，您需要了解您的数据，包括表、字段和内容，这就是 White Rabbit 工具的用处。White Rabbit 是一个软件工具，可帮助为纵向医疗数据库 ETL 到 OMOP CDM 的转化做准备。White Rabbit 将扫描您的数据并创建一个包含所有必要信息的报告，以开始设计 ETL。GitHub 上提供了所有源代码、安装说明以及使用手册的链接。

范围和目的

White Rabbit 的主要功能是执行源数据的扫描，提供有关表、字段和字段值的详细信息。源数据可以在逗号分隔的文本文件中，也可以在数据（MySQL，SQL Server，Oracle，PostgreSQL，Microsoft APS，Microsoft Access，Amazon RedShift）。White Rabbit 与标准数据分析工具的不同之处在于，它试图防止在生成的输出数据文件中显示个人身份信息（PII）数据。

流程概述

使用该软件扫描源数据的典型顺序：

1.设置工作文件夹，运行结果将导出到位于本地台式计算机上的该文件夹内。

2.连接到源数据库或 CSV 文本文件并测试连接。

3.选择要扫描的表并进行扫描。

4. White Rabbit 创建有关源数据的导出信息。

设置工作文件夹

下载并安装 White Rabbit 应用程序后，您需要做的第一件事是设置一个工作文件夹。 White Rabbit 创建的所有文件都将导出到此本地文件夹。点击图 6.1 中所示的”Pick Folder”按钮在您要扫描文档所在的本地环境中选择文档。

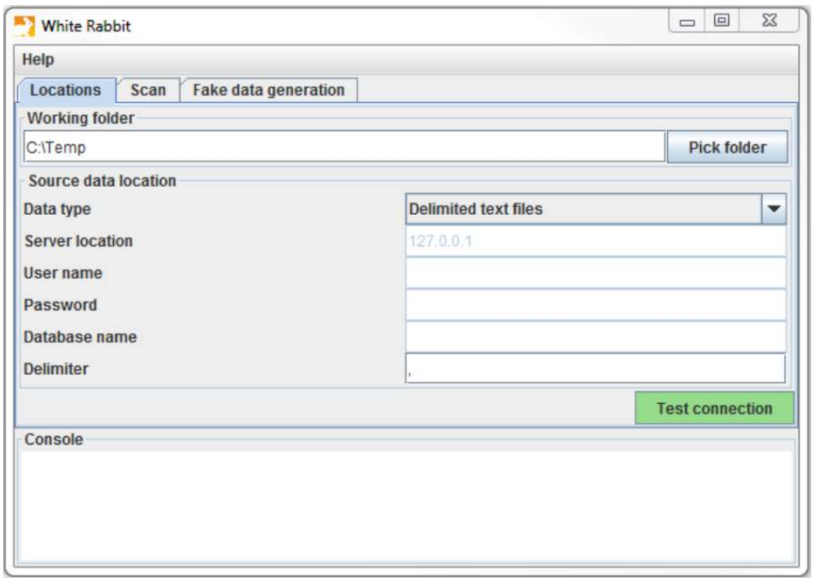


图 6.1：“选择文件夹”按钮允许为 White Rabbit 应用程序指定工作文件夹

连接到数据库

White Rabbit 支持带分隔符的文本文件和各种数据库平台。将鼠标悬停在各个字段上以获得所需内容的描述。您可以在手册中找到更多详细信息。

扫描数据库中的表

连接到数据库后，您可以扫描其中包含的表。扫描会生成一个包含有关源数据信息的报告，这些报告可用于帮助设计 ETL。使用图 6.2 中所示的“Scan”选项卡，可以通过单击“Add”（Ctrl +鼠标）在所选源数据库中选择单个表，也可以通过单击“Add all in DB”自动选择数据库中的所有表。

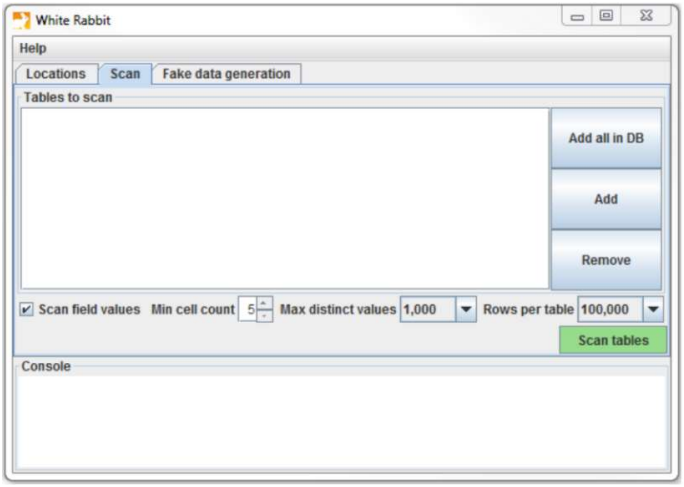


图 6.2：“White RabbitScan”选项卡

扫描还有一些设置选项：

•选择“Scan field values（扫描字段值）”会告诉 White Rabbit 您想查阅在扫描列中出现了哪些值。

•扫描字段值时，“Min cell count（最小单元格计数）”在默认情况下，此值设置为 5，表示源数据中出现少于 5 次的值将不会出现在报告中。单个数据集可能有自己的规则来确定这个最小单元格数可以是多少。

•扫描字段值时，可以选择“Rows per table（每表行数）”。默认情况下，White Rabbit 将扫描表中的 100,000 个随机选择的行。

完成所有设置后，按“Scan tables（扫描表格）”按钮。完成后，报告将被写入工作文件夹。

解读扫描报告

扫描完成后，将在所选文件夹中生成一个 Excel 文件，除了概览选项卡（Tab）外，扫描的表都提供了一个选项卡。 概览选项卡列出了所有扫描的表，表中的每个字段、字段的数据类型、字段的最大长度、表中的行数、扫描的行数以及找到每个频率为空的字段。图 6.3.显示示例概览标签。

	A	B	C	D	E	F	G
1	Table	Field	Type	Max length	N rows	N rows checked	Fraction empty
2	dbo.allergies	start	date	10	3184	3184	0
3	dbo.allergies	stop	date	10	3184	3184	0.725188442
4	dbo.allergies	patient	varchar	36	3184	3184	0
5	dbo.allergies	encounter	varchar	36	3184	3184	0
6	dbo.allergies	code	varchar	9	3184	3184	0
7	dbo.allergies	description	varchar	24	3184	3184	0
8							
9	dbo.careplans	id	varchar	36	30199	30199	0
10	dbo.careplans	start	date	10	30199	30199	0
11	dbo.careplans	stop	date	10	30199	30199	0.057849598
12	dbo.careplans	patient	varchar	36	30199	30199	0
13	dbo.careplans	encounter	varchar	36	30199	30199	0
14	dbo.careplans	code	varchar	15	30199	30199	0
15	dbo.careplans	description	varchar	62	30199	30199	0
16	dbo.careplans	reasoncode	varchar	9	30199	30199	0.050796384
17	dbo.careplans	reasondescription	varchar	56	30199	30199	0.050796384
18							

Overview | dbo.allergies | dbo.carepla ...

图 6.3：扫描报告中的示例概述选项卡

每个表的选项卡显示每个字段、字段值以及值的频率。源表中的每个列将在 Excel 中生成两列。一栏将列出所有数据非重复值（distinct values），该值出现次数大于“Min cell count（最小单元格计

数)”在扫描时设置的值。如果数据值列表被截断，则列表中的最后一个值将为“List truncated（列表被截断)”，这表示源列中，有一个或多个额外的唯一值出现次数小于“Min cell count（最小单元格计数)”中输入的数字。每个不重复值的旁边是第二个包含频率（该值在样本中出现的次数）的列，这两个列（不同的值和频率）将对工作簿中进行分析的表中的所有源列重复。

	A	B
1	Sex	Frequency
2	2	61491
3	1	35401
4	List truncated...	

图 6.4：单个列的示例值

该报告通过突出显示存在的内容，显示了理解源数据的强大能力。例如，如果图 6.4 中显示结果返回，在扫描的表中的“Sex”列上，我们可以看到有两个公共值（1 和 2）分别出现 61,491 和 35,401 次。White Rabbit 不会将 1 定义为男性，将 2 定义为女性，数据持有者通常需要定义源系统特有的源代码。但是，这两个值（1&2）并不是数据中出现的唯一值，因为我们看到此列表被截断了。其他值以非常低的频率出现（由“最小单元格计数”定义)，通常表示不正确或高度可疑的值。生成 ETL 时，我们不仅应该计划处理高频性别概念 1 和 2，还应该计划处理此列中存在的其他低频值。例如，如果那些较低频率的性别为“NULL”，我们希望确保 ETL 可以处理该数据并知道在这种情况下该怎么做。

6.2.2 Rabbit-in-a-Hat

借助“White Rabbit”扫描，我们可以清晰地看到源数据，也可以知道 CDM 的完整规范。现在我们需要定义从源数据到 CDM 的逻辑，此设计活动需要对源数据和 CDM 都有透彻的了解。White Rabbit 软件随附的 Rabbit-in-a-Hat 工具是专为支持这些领域的专家团队而设计的。在典型的环境中，ETL 设计团队会坐在一个房间里，而 Rat-in-a-Hat 投影在屏幕上。在第一轮中，团队可以共同决定表到表的映射，然后设计字段到字段的映射的同时定义将数值转换的逻辑。

范围和目的

Rabbit-in-a-Hat 设计用于读取和显示 White Rabbit 扫描文档。White Rabbit 生成关于源数据的信息，Rabbit-in-a-Hat 使用这些信息并通过图形用户界面允许用户将源数据连接到 CDM 中的表和列。Rabbit-in-a-Hat 为 ETL 过程生成文档，但不生成创建 ETL 的代码。

过程概述

使用此软件生成 ETL 文档的一般顺序：

1. White Rabbit 的扫描结果完成。

2.打开扫描结果，界面显示源表和 CDM 表。

3.将源表连接到 CDM 表，其中源表提供了对应的 CDM 表的信息。

4.根据每个源表到 CDM 表的连接，进一步定义源列到 CDM 列详细信息的连接。

5.保存 Rabbit-in-a-Hat 的工作并导出到 MS Word 文档。

编写 ETL 逻辑

在 Rabbit-in-a-Hat 中打开 White Rabbit 扫描报告后，您就可以开始设计和编写如何将源数据转

换为 OMOP CDM 的逻辑。例如，以下几节将介绍，Synthea 数据库中的某些表在转换期间的显示内容。

ETL 的一般流程

由于 CDM 是一个以个人为中心的模型，所以最好首先映射 PERSON 表。每个临床事件表（CONDITION_OCCURRENCE，DRUG_EXPOSURE，PROCEDURE_OCCURRENCE 等）都通过 PERSON_ID 引用 PERSON 表，因此首先确定 PERSON 表的逻辑将使以后的工作变得更加容易。在 PERSON 表完成之后，根据经验接下来是转换 OBSERVATION_PERIOD 表。CDM 数据库中的每个人都应至少有一个 OBSERVATION_PERIOD，一个人的大多数事件通常都在此时间范围内。一旦完成了 PERSON 和 OBSERVATION_PERIOD 表，通常接下来是 PROVIDER，CARE_SITE 和 LOCATION 之类的维表。在临床事件表之前应制定的最终表逻辑为 VISIT_OCCURRENCE，这是整个 ETL 中最复杂的逻辑，也是最关键的逻辑之一，因为在个体的病程中的大多数事件发生于就诊期间。一旦这些表完成，您就可以选择映射哪个 CDM 表，以及按哪个顺序映射。

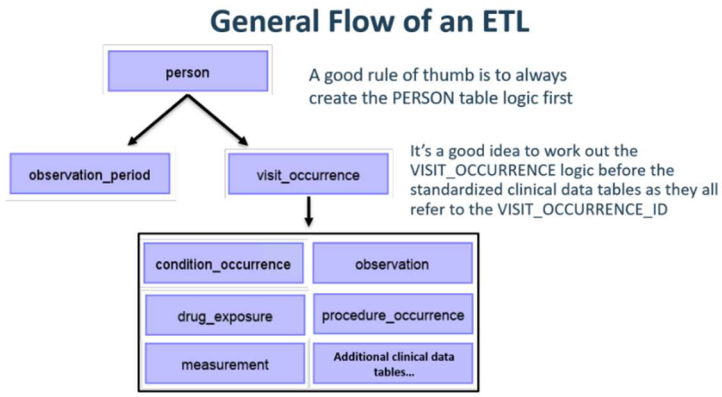


图 6.5：ETL 的一般流程以及首先要映射的表

通常，在 CDM 转换期间，您需要在中间表中进行准备，这是为了给事件分配正确的 VISIT_OCCURRENCE_ID，或者将源代码映射到标准概念（动态执行此步骤通常非常慢）。我们允许和鼓励中间表的使用，但并不鼓励转换完成后继续依赖这些中间表。

映射示例：Person 表

Synthea 数据结构在“患者（patient）”表中包含 20 列，但并非所有列都需要填充 PERSON 表，如图 6.6 所示，这很常见，不需引起警惕。比如 Synthea 患者表中在 CDM PERSON 表中使用的数据有附加标识符，例如患者姓名，驾照号码和护照号码。

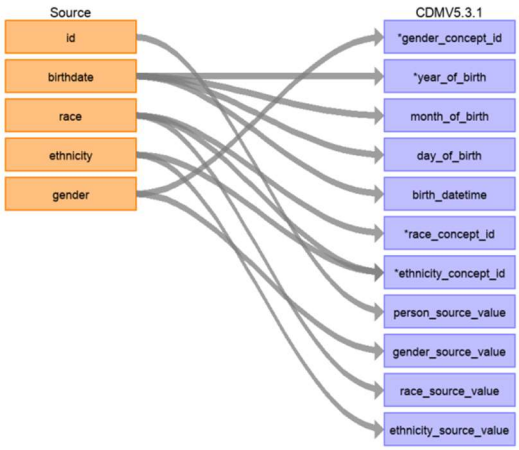


图 6.6：将 Synthea 患者表映射到 CDM PERSON 表

下表 6.1 显示了将 Synthea 患者表转换为 CDM PERSON 表的逻辑。“目标字段”显示了 CDM 数据要映射到的位置。“源字段”突出显示源表中的列（在本例中为患者），该列将用于填充 CDM 列。最后，“逻辑和注释”列提供了相关逻辑解释。

表6.1：将Synthea Patients表转换为CDM PERSON表的ETL逻辑

目标字段	源字段	逻辑与注释
PERSON_ID		自动生成。 PERSON_ID将在实施时生成。 这是因为来自源的id值是varchar值，而 PERSON_ID是整数。 来自源的id字段设置为PERSON_SOURCE_VALUE，以保留该值并在必要时允许进行错误检查。
GENDER_CONCEPT_ID	性别	如果性别='M'，则将GENDER_CONCEPT_ID设置为8507，而当性别='F'，则将其设置为8532。删除性别缺失或未知性别的行。选择这两个概念是因为它们是性别领域中仅有的两个标准概念。丢弃性别未知的患者的选择通常是基于站点的，但建议将其删除，因为没有性别的患者会被排除在分析之外。
YEAR_OF_BIRTH	出生日期	从出生日期中提取年份
MONTH_OF_BIRTH	出生日期	从出生日期中提取月份
DAY_OF_BIRTH	出生日期	从出生日期中提取出生日
BIRTH_DATETIME	出生日期	以午夜为时间00:00:00。此例中，来源没有提供出生时间，因此选择将其设置为午夜
RACE_CONCEPT_ID	种族	当race ='WHITE'设置为8527，当race ='BLACK'设置为8516，当race ='ASIAN'设置为8515，否则设置为0。选择这些概念是因为它们是属于的标准概念 与来源中的种族类别最接近的种族域。

有关如何将 Synthea 数据集映射到 CDM 的更多示例，请参见完整的规范文档。

6.3 步骤 2：创建代码映射

越来越多的数据源的代码表（简称源代码）被持续添加到 OMOP 词汇表中，这意味着要转换为

CDM 数据中的编码系统可能已经被包含和映射。检查 OMOP 词汇表中的 VOCABULARY 表，查看其中包含哪些词汇表，要将非标准源代码（例如 ICD-10CM）到标准概念（例如 SNOMED CT）的映射提取出来，我们可以使用 CONCEPT_RELATIONSHIP 表中具有 RELATIONSHIP_ID=“Maps To”的记录。例如，要查找 ICD-10CM 代码“ I21”（“急性心肌梗塞”）的标准概念 ID，可以使用以下 SQL:

```
SELECT concept_id_2 standard_concept_id
FROM concept_relationship
INNER JOIN concept source_concept
      ON concept_id = concept_id_1
WHERE concept_code = 'I21'
      AND vocabulary_id = 'ICD10CM'
      AND relationship_id = 'Maps to';
```

但是，有时源数据使用不在词汇表中的编码系统，在这种情况下，必须从源编码系统创建到标准概念的映射。术语映射可能是一项艰巨的任务，特别是当源代码系统中有许多编码系统时，下述做法可使该任务相对容易：

•关注最常用的编码。不会使用或不经常使用的编码不值得进行映射，因为在真正的研究中永远不会使用它。

•尽可能利用现有信息。例如，许多国家药品编码系统已被映射到 ATC。尽管 ATC 对于许多用途而言不够详细，但是 ATC 和 RxNorm 之间的概念关系可以用来很好地猜测正确的 RxNorm 代码是什么。

•使用 Usagi。

6.3.1 Usagi

Usagi 是辅助手动创建代码映射的工具。它可以基于代码描述的文本相似性来建议映射。如果源代码仅以外语提供，我们发现谷歌翻译通常会将术语很好的翻译为英语。如果自动建议不正确，Usagi 允许用户搜索适当的目标概念。最后，用户可以指示哪些映射被批准在 ETL 中使用。Usagi 可在 GitHub 找到。

范围和目的

需要映射的源代码已加载到 Usagi 中（如果代码不是英语，则需要附加翻译列）。术语相似性方法用于将源代码连接到词汇表概念。但这些代码连接需要手动评审，并且 Usagi 提供了一个界面来简化此操作。Usagi 仅会推荐在词汇表中标记为“标准概念”的概念。

过程概述

使用该软件的典型流程是：

1.从您的源系统（“源代码”）中加载要映射到词汇表概念的代码。

2.Usagi 将使用术语相似性方法将源代码映射到词汇表概念。

3.利用 Usagi 界面进行检查，并在需要时改进建议的映射。推荐选择具有编码系统和医学术语经验的个人来进行此审核。

4.将映射导出到词汇表的 SOURCE_TO_CONCEPT_MAP。

将源代码导入 Usagi

将源代码从源系统导出到 CSV 或 Excel（.xlsx ）文件。该字段至少应包含源代码及其英语描述的列，但是也可包括有关代码的其他信息（例如剂量单位，或者源语言的描述）。除了有关源代码的信息之外，还应优先考虑代码的频率，因为这可帮助在代码优化过程中进行优先级排定（例如，您可以有 1000 个源代码，但在系统中仅真正使用了 100 个）。如果任何源代码信息需要翻译成英文，请使用“谷歌翻译”。

注：源代码摘录应按领域（即药物、操作、疾病、观测）分类，而不应集中在一个大文件中。

从“文件–>导入代码”菜单，将源代码加载到 Usagi 中。从这开始，将显示“导入代码...”，如图 6.7 所示。在此图中，源代码术语是荷兰语，也被翻译成英语。Usagi 将利用其英语翻译来映射到标准词汇表。

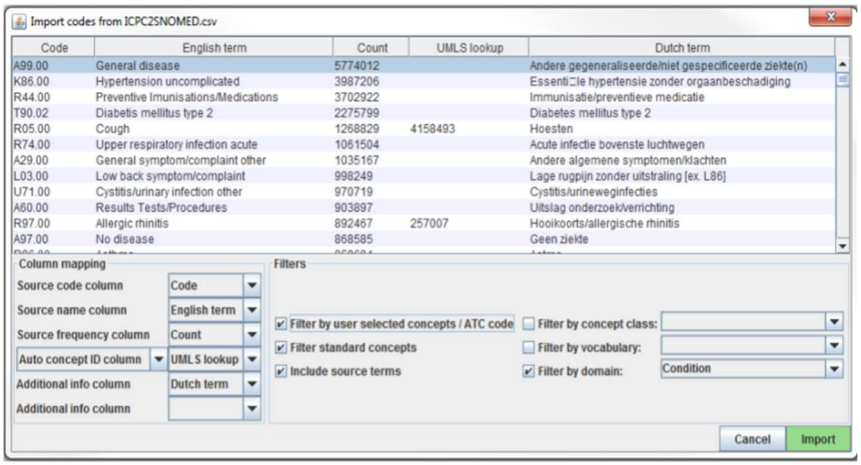


图 6.7：Usagi 源代码输入屏幕。

您可以在“列映射”部分（左下方）为 Usagi 定义如何使用导入的表。如果将鼠标悬停在下拉菜单上，将出现一个弹出窗口，定义每列。Usagi 不会将“其他信息”列用作将源代码与词汇概念代码相关联的信息；但是，此附加信息可能有助于个人检查源代码映射，因此应该包括在内。

最后，在“过滤器”部分（右下），您可以在映射时为 Usagi 设置一些限制。例如，在图 6.7 中，用户仅将源代码映射到条件域中的概念。默认情况下，Usagi 仅映射到标准概念，但是如果关闭了“过滤标准概念”选项，Usagi 还将考虑分类概念。将鼠标悬停在不同的过滤器上以获取有关过滤器的其他信息。

一种特殊的过滤器是“通过自动选择的概念/ ATC 代码过滤”。如果存在可用于限制搜索的信息，则可以通过在“自动概念 ID”指示列中，提供 CONCEPT_ID 或 ATC 代码的列表（以分号分隔的方式）来进行限制。例如，对于药物，可能已经为每种药物分配了 ATC 代码。即使 ATC 代码不能唯一地标识单个 RxNorm 药品代码，它也确实有助于将搜索空间限制为仅属于词汇表中 ATC 代码下的那些概念。要使用 ATC 代码，请按照下列步骤操作：

1. 在“列映射”部分中，从“自动概念 ID 列”切换到“ ATC 列”

2. 在“列映射”部分中，将包含 ATC 代码的列选择为 “ ATC 列”。

3. 打开“过滤器” 部分中的“按用户选择的概念/ ATC 代码进行过滤”。

您还可以使用 ATC 代码以外的其他信息源进行限制。在上图所示的示例中，我们使用了从 UMLS 派生的部分映射来限制 Usagi 搜索。在这种情况下，我们将需要使用“自动概念 ID 列”。

完成所有设置后，单击“导入”按钮导入文件。由于正在运行术语相似性算法来映射源代码，因此文件导入将需要几分钟。

审核词源码-词汇概念映射

导入源代码输入文件后，映射过程即开始。在图 6.8 中，您会看到 Usagi 屏幕由 3 个主要部分组成：概述表，所选的映射部分和执行搜索的位置。请注意，在任何表格中，您都可以右键单击以选择显示或隐藏的列，以减少视觉复杂性。

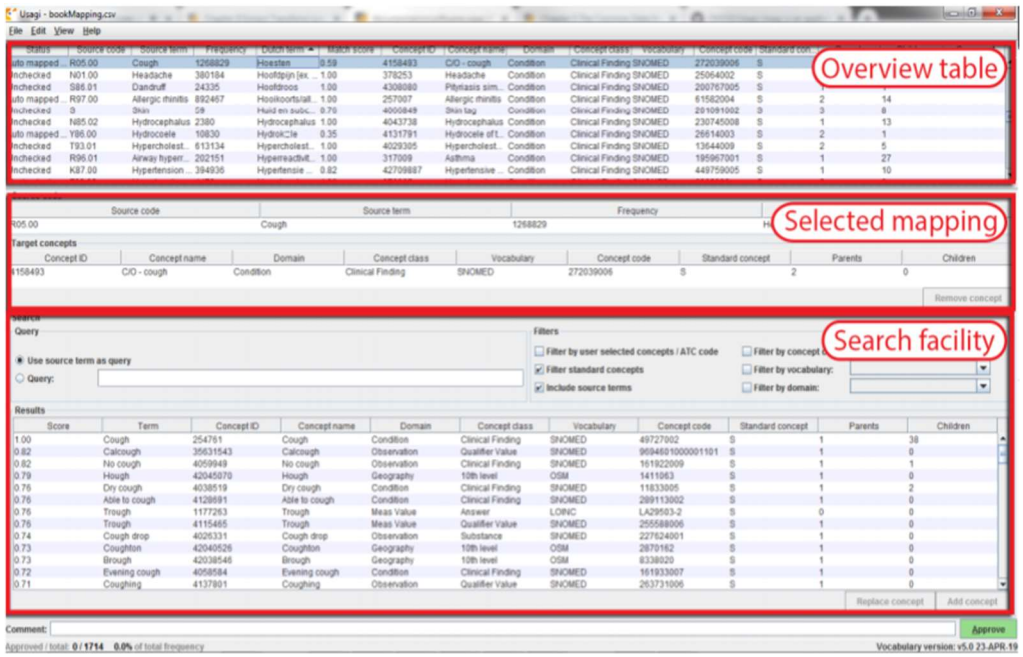


图 6.8：Usagi 源代码输入屏幕。

批准建议的映射

“概述表”显示了源代码到概念的当前映射。导入源代码后，此映射将包含基于术语相似性和任何搜索选项所自动生成的建议映射。在图 6.8 的示例中，荷兰语条件代码的英文名称被映射到“条件域”中的标准概念，因为用户将搜索限制在该域中。Usagi 将源代码描述与概念名称和同义词进行了比较，以找到最佳匹配。由于用户选择了 “包括源术语”，Usagi 还考虑了词汇表中映射到特定概念的所有源概念的名称和同义词。如果 Usagi 无法进行映射，它将映射到 CONCEPT_ID = 0。

建议由具有编码系统经验的人来帮助将源代码映射到其关联的标准词汇表。该人员将基于“概述表”中的代码，逐一接受 Usagi 建议的映射或选择新的映射。例如在图 6.8zh 中，我们看到荷兰 “ Hoesten ” 被翻译成英语 “咳嗽”。Usagi 使用 “咳嗽” 并将其映射到 “ 4158493-C / O-咳嗽” 的词汇概念。与此匹配对相关的匹配分数为 0.58（匹配分数通常为 0 到 1，其中 1 为可信匹配），得分

0.58 表示 Usagi 不太确定将荷兰语代码映射到 SNOMED CT 的程度。 我们看来，在这种情况下，我们可以接受这种映射，可以通过点击屏幕右下角绿色的“批准”按钮来批准它。

搜索新的映射

在某些情况下，Usagi 将建议一个映射表，而用户将被迫尝试寻找更好的映射表或将其设置为无概念（CONCEPT_ID = 0）。在图 6.8 所示的示例中，我们看到荷兰语术语 “ Hoesten ”，该术语已翻译为“咳嗽”。Usagi 的建议受到我们从 UMLS 自动导出的映射中确定的概念的限制，而其可能不是最佳的。在搜索工具中，我们可以使用实际术语本身或搜索框查询来搜索其他概念。

使用手动搜索框时，应记住 Usagi 使用模糊搜索，并且不支持结构化搜索查询，例如，不支持 AND 和 OR 之类的布尔运算符。

继续我们的示例，假设我们使用搜索词 “咳嗽” 来查看是否可以找到更好的映射。在搜索工具的 “查询” 部分的右侧，有一个 “过滤器” 部分，它提供了一些选项，可用于在搜索搜索词时从词汇表中修剪结果。在这种情况下，我们知道我们只希望找到标准概念，并且允许基于映射到这些标准概念的词汇中源概念的名称和同义词来找到概念。

应用这些搜索条件时，我们发现 “ 254761-咳嗽”，并认为这可能是映射到荷兰语代码的适当词汇概念。为此，我们可以单击 “替换概念”，您将在 “选定的源代码” 部分中看到更新，然后单击 “批准”。此外 “添加概念”，它允许多个标准化词汇概念映射到一个源代码（例如，某些源代码可能将多种疾病捆绑在一起，而标准化词汇可能没有）。

概念信息

在寻找适当的概念来映射时，重要的是要考虑概念的 “社会特性”。概念的含义可能部分取决于其在层次结构中的位置，有时词汇中的 “孤立概念” 几乎没有或没有层次关系，因此不适合作为目标概念。Usagi 经常会报告一个概念所拥有的父级及子级概念的数量，还可以通过按 “ALT + C” 或在顶部菜单栏中选择 “查看–>概念信息” 来显示更多信息。

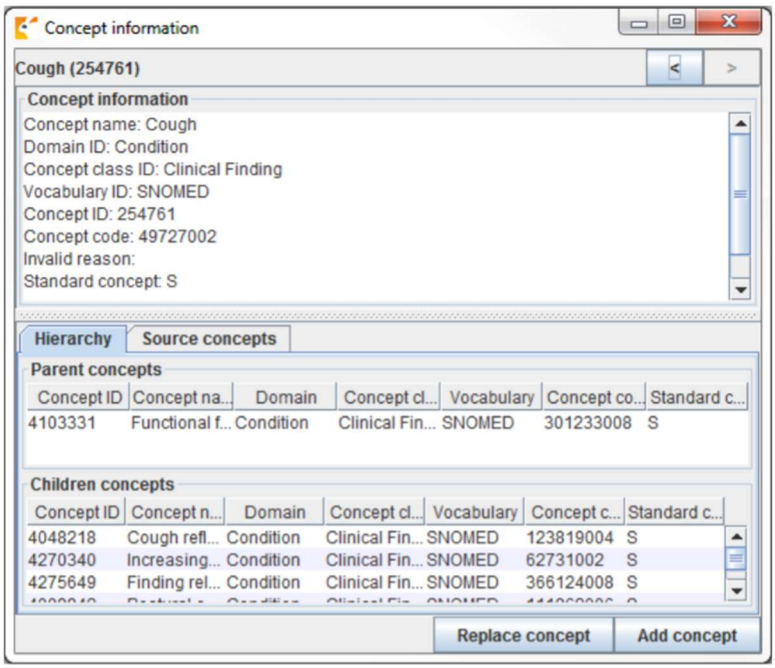


图 6.9：Usagi 概念信息面板。

图 6.9 显示了概念信息面板。它显示有关概念的一般信息，以及它的父级，子级和映射到该概念的其他源代码。用户可以使用此面板浏览层次结构，并可以潜在的选择其他目标概念。继续逐个代码地执行此过程，直到检查完所有代码。在屏幕顶部的源代码列表中，通过选择列标题可以对代码进行排序。通常，我们建议从最高频率到最低频率顺序排列。在屏幕的左下方，您可以看到已批准映射的代码数，以及与之对应的代码出现次数。

可以在映射中添加注释，这些注释可用于记录做出决策的原因。

最佳实践

- 使用具有编码经验的人员。
- 通过单击列名，可以在“概述表”中对列进行排序。对“匹配分数”进行排序可能很有价值；首先查看 Usagi 最可信的代码，这样可能会很快排除大量代码。同样，对“频率”进行排序也很有价值，在频繁代码上比非频繁代码花费更多的精力也很重要。
- 将一些代码映射为 CONCEPT_ID = 0 是可以的，因为有些代码可能不值得找到一个好的映射，而另一些代码可能只是缺少适当的映射。
- 重要的是要考虑一个概念的背景环境，特别是它的父级和子级。

导出创建的 Usagi 映射表

在 USAGI 中创建映射表后，进一步使用它的最佳方法是将其导出并附加到 Vocabulary SOURCE_TO_CONCEPT_MAP 表中。

要导出映射，请转到“文件–>导出”source_to_concept_map 。将弹出一个窗口，询问您要使用哪个 SOURCE_VOCABULARY_ID，并输入一个简短的标识符。Usagi 将使用此标识符，作为 SOURCE_VOCABULARY_ID。这将允许您在 SOURCE_TO_CONCEPT_MAP 表中标识您的特定映射。

选择 SOURCE_VOCABULARY_ID 之后，为导出的 CSV 命名并保存。导出的 CSV 结构即 SOURCE_TO_CONCEPT_MAP 表的结构。该映射可以附加到“词汇表”的 SOURCE_TO_CONCEPT_MAP 表中。此外，也可将单行附加到“词汇表”以定义您在上述步骤中定义的 SOURCE_VOCABULARY_ID。最后，需要注意的是，只有状态为“已批准”的映射才会导出到 CSV 文件中。映射需要在 USAGI 中完成才能导出。

更新 Usagi 映射

映射通常不是一次性的工作。随着数据的更新，也许会添加新的源代码，并且词汇表会定期更新，也许需要更新映射。

当更新源代码集时，可以遵循以下步骤：

1．导入新的源代码文件

2．选择“文件->应用先前的映射“，然后选择旧的 Usagi 映射文件

3．标识未从旧映射中继承批准的映射的代码，并进行映射。

词汇表更新后，请按照下列步骤操作：

1．从 Athena 数据库下载新的词汇文件

2．重建 Usagi 索引（帮助–>重建索引）

3．打开映射文件基于新词汇版本的概念，确定哪些不再是标准概念的代码，并找到更合适的目标概念。

6.4 步骤 3: ETL 实施

一旦设计和代码映射完成，就可在软件中实现 ETL 过程。 在设计 ETL 时，我们建议熟悉源代码和 CDM 的人员一起完成任务。同样，在实施 ETL 时，人员最好具有处理数据（尤其是大数据）经验，并且具有实施 ETL 经验。这就意味着要与您的直属团队之外的人员一起工作，或雇用技术顾问来进行实施。同样重要的是，这不是一次性投入。项目执行过程中也要有人或团队投入时间来维护和运行 ETL（在 6.7 章节中有更详尽描述）。

实施通常因站点而异，并且在很大程度上取决于诸多因素，包括基础架构，数据库大小，ETL 的复杂性以及可用的技术专家。由于考虑因素众多，所以 OHDSI 社区并未就如何最好地实施 ETL 提出正式建议。已有讨论组基于简化 SQL 生成器，SAS，C＃，Java 和 Kettle 进行工作。所有这些都有其优缺点，但是如果站点没有人熟悉这些技术，就将无法使用。

一些不同 ETL 的例子 （根据其复杂度排序）：

- ETL-Synthea - SQL 生成器，用于转换 Synthea 数据库 https://github.com/OHDSI/etl-synthea
- ETL-CDMBuilder - .NET 应用，设计用于转换多数据库 https://github.com/OHDSI/etl-cdmbuilder
- ETL-LambdaBuilder - 基于亚马逊云 Lambda 功能的生成器 https://github.com/OHDSI/etl-lambdabuilder

应该指出的是，在经过几次独立尝试之后，我们已经放弃开发“终极的”用户友好型 ETL 工具。通常情况下，此类工具对于 80%的 ETL 都适用，但是对于其余 20%的 ETL，则需要编写一些特定于源数据库的底层代码。

一旦技术人员准备开始实施，就应该与他们共享 ETL 设计文档。文档中应该有足够的信息供他们启动工作。不仅如此，也要保证负责实施的开发人员在实施过程中能够与 ETL 设计人员联系并提出问题。设计者可能清楚其逻辑，但是对于不熟悉数据和 CDM 的实施者来说就可能不太清晰了。因此，实施阶段需保持团队合作精神。实施者和设计者可共同核对厘清 CDM 创建和测试的过程，直到两组都同意所有逻辑均已被正确执行。

6.5 步骤 4: 质量控制

对于提取、转换、加载过程，质量控制是迭代的。典型的模式是编写逻辑->实施逻辑->测试逻辑->修复/写入逻辑。有多种测试 CDM 的方法，但以下建议的步骤，是整个社区在多年 ETL 实施中建立起来的。

- 审核 ETL 设计文档，计算机代码和代码映射。任何人都可能犯错，因此必须要由其他人员来审核已完成的工作。
- 计算机代码中最大的问题是，原始数据中的源代码如何映射到标准概念。映射可能会变得棘手，尤其是在涉及特定日期的代码（例如 NDC）时。仔细检查所有完成映射的区域，以确保将正确的源词汇表转换为正确的概念 ID。
- 手动比较源数据和目标数据中有关样本人员的所有信息。
- 完整浏览单个个体的数据（最好是拥有大量唯一记录的人）会很有帮助。如果 CDM 中的数

据不是基于预期逻辑展现的，通过对特定样本进行跟踪会突显一些问题。

- 比较源数据和目标数据中的总计数。
- 根据您选择解决特定问题的方式不同，计数可能会差异。例如，一些协作者选择删除性别为 NULL 的个体，因为无论如何这些人都不会被包括在分析中。同样也可能存在，CDM 中构造的就诊次数与原始数据不同的情况。因此，在比较源数据和 CDM 数据之间的总计数时，请务必考虑并预估这些差异。
- 将已研究的源数据复制到 CDM 上
- 这是了解源数据和 CDM 版本之间存在任何主要差异的好方法，尽管会花费更多时间。
- 创建单元测试，意味着复制需要 ETL 处理的某个源数据中的模式。例如，如果您的 ETL 指定应删除没有性别信息的患者，请创建一个没有性别信息的样本的单元测试，并核对生成器的处理过程。
- 在评估 ETL 转换的质量和准确性时，单元测试非常方便。通常创建一个更小的数据集，以模仿要转换的源数据的结构。数据集中的每个样本或记录都应测试 ETL 文档中编写的特定逻辑。使用此方法，很容易追溯问题并识别错误的逻辑。小数据集还使得计算机代码可以快速地执行，从而可以实现更快的迭代和错误识别。

这些是从 ETL 角度进行质量控制的高级方法。有关更多 OHDSI 正在进行的数据质控工作的详细信息，请参见本书第 15 章。

6.6 ETL 约定及 THEMIS

随着越来越多的机构将数据转换为 CDM，很明显需要明确相应的约定。例如，在个人记录缺少出生年份的情况下，ETL 应该怎么做？CDM 的目标是使医疗保健数据标准化，但是，如果每个小组对数据特定情况的处理方式有所不同，则将更加难以在整个网络上系统地使用数据。

OHDSI 社区已做出相关约定，以提高 CDM 之间的一致性。可以在 CDM Wiki 查找到 OHDSI 社区已批准且定义好的约定。每个 CDM 表都有其自己的约定集，在设计 ETL 时可以参考这些约定。例如，允许样本缺失出生月份或日期，但是如果他们缺少出生年份，则应将其丢弃。在设计 ETL 时，请参考特定约定以保证和社区一致的设计决策。

虽然不可能记录所有的数据场景以及它们发生时的处理方法，但 OHDSI 工作组试图记录常见的应用场景。THEMIS 由 OHDSI 社区成员组成，他们收集约定，对约定进行阐释，共享约定并征集意见，在 CDM Wiki 中记录最终的约定。Themis 源于古希腊的泰坦神，它象征着神圣的秩序、公平、法律、自然法则和风俗习惯，正符合 OHDSI THEMIS 工作组的特点。在执行 ETL 时，如果不确定如何处理某些场景，THEMIS 建议在 OHDSI 论坛上讨论有关该场景问题。如果您存在问题，很可能是社区中的其他人也有相同问题。 THEMIS 基于这些讨论、工作组会议和面对面讨论来帮助确定哪些公约需要被记录。

6.7 CDM 及 ETL 维护

设计 ETL，创建映射，实现 ETL 并制定质控措施，这并非易事。然而，工作并不止步于此。建立第一个 CDM 之后，ETL 的周期性维护将是一个持续的过程。一些常见的触发维护的事件包括：源数据更改、ETL 错误、新发布的 OMOP 词汇表、或 CDM 自身的更改或更新。如果发生这些事件，则可能需

要更新以下内容：ETL 文档，运行 ETL 的软件程序，以及测试用例和质量控制。

通常，医疗健康数据源永远都在变化。新的可用数据可能随时出现（例如，数据中可能存在新列）。新的患者场景可能会突然出现在数据中（例如，一名新患者在出生前有死亡记录）。对数据的理解可能得到提升（例如，基于处理索赔的方式，一些住院婴儿的出生记录会作为门诊病人出现）。并非源数据中的所有更改都会触发 ETL 处理的更改，但是至少需要解决那些破坏 ETL 处理的更改。

如果发现错误，则应予以解决。bug 的起源和重要性各有不同。例如，假设在 COST 表中，列的费用已四舍五入为一个整数（例如，源数据为$ 3.82，而在 CDM 中为$ 4.00）。如果使用这些数据的研究人员，主要是对患者的药物暴露和状况进行研究的，那么此类漏洞的重要性就很小，可以将来解决。如果使用数据的主要研究人员还包括卫生经济学家，这将是一个关键漏洞，需要立即解决。

就像我们的源数据一样，OMOP 词汇也在不断变化。实际上，随着词汇表的更新，词汇表可以在某些月份内发布多个版本。每个 CDM 在特定版本的词汇表上运行，而在更新后的词汇表上运行时，可能会导致源代码在标准词汇表中的映射方式发生变化。通常，词汇表之间的差异很小，因此无需在每次发布新词汇表时都构建新的 CDM。但是，最佳实践是每年更新一次或两次词汇表，同时需要重新构建 CDM。因词汇表更新而导致 ETL 代码需要更新的情况很少发生。

最后一类触发 CDM 或 ETL 维护的事件是通用数据模型本身发生了更新。随着 OHDSI 的发展和新数据需求的发掘，可能会将更多数据存入 CDM 中。这就意味着，以前未出现的一些数据，将在新的 CDM 版本中占据位置。更改现有 CDM 结构的可能性较小，但可能存在。例如，由于 CDM 中，原 DATE 字段上采用 DATETIME 格式，这可能会导致 ETL 处理产生错误。 CDM 版本并不经常发布，各个站点可以在迁移时选择合适版本。

6.8 ETL 最终思考

ETL 流程很难掌握，原因有很多，其中最重要的一个原因是，我们需要处理源数据来源各异，从而无法创建普适性的解决方案。尽管如此，我们也积累了以下的经验。

- 80/20 规则。尽量不要花费太多时间手动将源代码映射到概念集。理想情况下，映射那些覆盖了绝大部分数据的源代码。这应该足以保证工作的启动，后续可以根据用例解决剩余的代码映射。
- 剔除未达到研究质量的数据，不要犹豫。这些记录到了分析阶段也通常会被排除在外，我们只是在 ETL 的阶段就执行了剔除而已。
- CDM 需要维护。完成了一次 ETL 并不意味着不需要对 CDM 进行维护。原始数据可能会更改，代码中可能有错误，可能有新词汇表或 CDM 更新，这些都有可能触发 CDM 的更新。在工作计划为 CDM 更新分配必要的资源，从而保证 ETL 始终是最新的。
- 要获得 OHDSI CDM 入门，执行数据库转换或运行分析工具的支持，请访问我们的实施者论坛（Implementers Forum）。

6.9 总结

有关如何处理ETL的流程已达成共识，包括：

- 数据专家和CDM专家共同设计ETL；
- 具有医学知识的人创建代码映射；
- 技术人员实施ETL；
- 所有步骤均需质量控制。

OHDSI 社区开发了工具来简化这些步骤，并且可以免费使用。有许多 ETL 示例和既成约定，可以用于指导。

6.10 练习

练习 6.1. 以适当的顺序实现 ETL 过程的步骤：

A）数据专家和 CDM 专家共同设计 ETL

B）技术人员实施 ETL

C）有医学知识的人创建代码映射

D）所有参与质量控制

练习 6.2. 使用您选择的 OHDSI 资源，在表 6.3（表缩写为空格）中发现 PERSON 记录的四个问题：

表 6.2：一个 PERSON 表

Column	Value
PERSON_ID	A123B456
GENDER_CONCEPT_ID	8532
YEAR_OF_BIRTH	NULL
MONTH_OF_BIRTH	NULL
DAY_OF_BIRTH	NULL
RACE_CONCEPT_ID	0
ETHNICITY_CONCEPT_ID	8527
PERSON_SOURCE_VALUE	A123B456
GENDER_SOURCE_VALUE	F
RACE_SOURCE_VALUE	WHITE
ETHNICITY_SOURCE_VALUE	NONE PROVIDED

练习 6.3. 我们尝试生成 VISIT_OCCURRENCE 记录。这是为 Synthea 编写的一些逻辑示例：按 PATIENT、START、END 以升序对数据进行排序。然后只要一行的结束到下一行的开始之间的时间小于<= 1 天，通过 PERSON_ID 就可以折叠声明的行。将每个合并的住院索赔视为一次住院访问，设置

为:

- MIN(START) 设置为 VISIT_START_DATE
- MAX(END) 设置为 VISIT_END_DATE
- “IP” 设置为 PLACE_OF_SERVICE_SOURCE_VALUE

如果您在源数据中看到如图 6.10 所示的一组访问，那么如何在 CDM 中显示期望生成的 VISIT_OCCURRENCE 记录?

Data Output | Explain | Messages | Notifications | Query History

	id character varying (1000)	start date	stop date	patient character varying (1000)	encounterclass character varying (1000)
1	12	2004-09-26	2004-09-27	11	inpatient
2	13	2004-09-27	2004-09-30	11	inpatient

图 6.10：示例源数据。

建议在附录 E.3.中找到答案

---

30. https://github.com/OHDSI/WhiteRabbit.
31. Synthea™ is a patient generator that aims to model real patients. Data are created based on parameters passed to the application.The structure of the data can be found here: https://github.com/synthetichealth/synthea/wiki
32. https://ohdsi.github.io/ETL-Synthea/
33. https://translate.google.com/
34. https://github.com/OHDSI/Usagi
35. https://github.com/OHDSI/CommonDataModel/wiki
36. https://github.com/OHDSI/Themis
37. http://forums.ohdsi.org/
38. https://forums.ohdsi.org/c/implementers

第七章　数据分析用例

章节负责人：大卫·麦迪根（David Madigan）

OHDSI 合作组织通常以理赔数据库或电子健康档案库的形式，从真实医疗数据中得到可靠的证据。OHDSI 重点关注的病例分析可分为三大类：

- 特征刻画
- 群体水平评估
- 患者水平预测

我们将在下面详细介绍这些内容。注意，对于所有用例，我们得到的证据都受限于数据本身的局限性。我们将在本书的“结论质量”一章中详细讨论这些局限（第 14-18 章）。

7.1 特征

刻画特征是用来回答以下问题：他们怎么了？

我们可以使用这些数据来回答有关某个队列或整个数据库中人员特征、医疗保健实施的问题，并研究这些事件如何随时间的变化。

这些数据可以回答以下这些类型的问题：

- 对于新诊断为心房纤颤的患者，医生给多少人开了华法林处方？
- 接受人工全髋关节置换术的患者的平均年龄是多少？
- 65 岁以上患者的肺炎发病率是多少？

有关特征的典型问题如下所示：

- 有多少患者...？
- 频率...？
- 患者的比例是...？
- 实验室的检测结果分布是什么...？
- 患有...（某些疾病）的患者的 HbA1c 水平是多少？
- 患者的实验数值是多少？
- ...（某些疾病）患者暴露时间的中位数是多少？
- 随着时间的推移，趋势是什么？
- 这些患者还使用哪些其他药物？
- 伴随疗法是什么？
- 我们是否有足够的...（某些疾病）案例？
- 研究 X...可行吗？
- ...（某些疾病）的人口统计特征是什么？
- ...（某些疾病）的危险因素是什么？　（如果确定特定的风险因素，可能是估计的，而不是预测的）
- ...（某些疾病）的预测因素是什么？

期望得到的结果是：

- 计数或百分比
- 均值
- 描述性统计量
- 发病率
- 患病率
- 队列
- 基于规则的表型
- 药品使用
- 疾病自然史
- 依从性
- 并发症
- 治疗途径
- 诊疗流程

7.2 群体水平评估

在一定程度上，数据可以支持对医疗保健干预措施影响的因果推断，从而回答关键问题：

有什么因果关系？

我们想了解通过因果关系，从而了解操作的后果。例如，如果我们决定接受某种治疗，那么这将如何改变我们将来的情况？

该数据可以回答类似以下的问题：

- 对于刚被确诊为房颤的患者，在治疗开始后的第一年内，使用华法林是否比使用达比加群引起更多的大出血？
- 二甲双胍对腹泻的影响是否随年龄而变化？

有关群体水平评估的典型问题如下所示：

- ...的作用是什么？
- 如果我进行...干预会怎样？
- 哪种治疗效果更好？
- X 作用在 Y 上的风险是什么？
- ...事件发生时间是几点？

期望得到的结果是:

- 相对风险
- 风险比
- 优势比
- 平均干预效应
- 因果关系
- 关联性
- 相关性
- 安全性检测

- 比较疗效

7.3 患者水平预测

根据数据库中收集的患者健康历史记录，我们可以对未来的健康事件做出患者水平预测，从而回答问题：我会怎样？

该数据可以回答类似以下的问题：

- 对于新诊断为重度抑郁症的特定患者，该患者在诊断后第一年自杀的可能性是多少？
- 对于刚被诊断为房颤的特定患者，在开始使用华法林治疗后的第一年，该患者发生缺血性中风的可能性是多少？
- 有关患者水平预测的典型问题如下所示：
- 该患者将有多少几率会……？
- 谁是…的潜在风险人群？

期望得到的结果是:

- 个体患病概率
- 预测模型
- 高/低风险人群
- 概率表型

群体水平评估和患者水平预测会在一定程度上重叠。例如，预测中的重要用例：如果处方中开了药物 A，可以预测某个特定患者的结局，而如果处方中开了药物 B，可以预测相同的结果。假设实际上医生只开出了其中一种药物（比如说药物 A），因此我们只能看到使用药物 A 后的结果。由于事实上并没有开药物 B，因此药物 B 的结果虽然可以预测，但却是“反事实”的，因为从未被观察到。这些预测任务都属于患者水平预测。但是，两个结果之间的差异（或比例）是单位级别的因果效应，应使用因果效应估计方法进行估计。

人们天生就有将预测模型错误地解释为因果模型的倾向。但是预测模型只能显示相关性，而不能显示因果关系。 例如，糖尿病药物的使用可能导致心肌梗塞（MI）的重要预测指标，因为糖尿病是 MI 的重要危险因素。但是，这并不意味着停止糖尿病药物可预防心梗！

7.4 高血压方面的应用案例

您是一位研究急性心肌梗塞和血管性水肿等一线治疗高血压方法的研究人员，您对 ACE 抑制剂单一疗法与噻嗪类利尿剂单一疗法的疗效感兴趣。根据 OHDSI 文献，您了解到您正在询问的是一个人群水平的效应估计问题，但首先，您需要做一些关于如何表征这种特殊目标疗法的功课。你了解到根据 OHDIS 上的文献要求，在你将要问一个群体层面的效应估计问题之前，但你首先做一些关于如何表征这一特指的目标疗法的功课。

7.4.1 特征问题

急性心肌梗塞是一种可能发生在高血压患者身上的心血管并发症，因此对高血压的有效治疗应降低急性心肌梗塞的风险。血管性水肿是 ACE 抑制剂已知的一种很少见但可能很严重的副作用。首先，您要针对暴露人群 (ACE 抑制剂的新使用者和噻嗪类利尿剂的新使用者) 建立队列 (请参阅第 10 章)。您需要进行特征分析（请参见第 11 章）以总结这些暴露人群的基本特征，包括人口统计学，合并病症和伴随用药。您要进行另外的特征分析，来估计在暴露人群中这些所选结果的发生率。对此您要问“急性心肌梗死多久发生一次？”，“在服用 ACE 抑制剂和噻嗪类利尿剂期间血管性水肿多久发生一次？”这些特征使我们能够评估进行人群水平估计研究的可行性、以及两个治疗组是否具有可比性，并确定可以预测患者选择哪种治疗方法的“风险因素”。

7.4.2 群体水平评估问题

群体水平的效应估计研究（请参阅第 12 章）是估计 ACE 抑制剂与噻嗪类药物用于 AMI 和血管性水肿的相对风险。对此，您可以通过研究诊断学和阴性对照进一步评估我们是否可以对它们的平均治疗效果做出可靠的估计。

7.4.3 患者水平预测问题

不论是否与患者有因果关系，您都想尝试确定哪些患者由最高风险患病。 这是一个患者水平预测问题（请参阅第 13 章）。对此您将建立一个预测模型，该模型用来评估：在刚使用 ACE 抑制剂的患者中，哪些患者在开始治疗后 1 年内发生急性心肌梗塞的风险最高。该模型使我们能够根据第一次接受 ACE 治疗患者的病史中观察到的事件，预测他们在接下来的一年中发生 AMI 的几率。

7.5 观察研究的局限性

有许多重要的医疗健康问题，OHDSI 数据库无法提供答案。其中包括：

干预治疗的因果效应与安慰剂对比。有时有可能分析出治疗与不治疗的因果效应,但不是和安慰剂治疗相比。

任何东西都与非处方药有关。

很多结果和其他变量都很少被记录下来。包括死亡率、行为结果、生活方式和社会经济状况。

患者倾向于在身体不适的时候才会接触医疗健康，因此衡量治疗的益处是很难的。

7.5.1 错误的数据

OHDSI 数据库中记录的临床数据可能会偏离临床实际。例如，即使患者从未经历过心肌梗死，患者的记录也可能包括心肌梗死代码。类似地，lab 值可能是错误的，或者由某一程序中的错误代码可能出现在数据库中。第 15 章和第 16 章讨论了其中的一些问题，良好做法旨在尽可能多地识别和纠正这些问题。然而，错误的数据不可避免地会在一定程度上持续存在，并可能损害后续分析的有效性。大量的文献更关注于对统计推断的调整，以解释数据中的错误-例如，见 Fuller（2009）。

7.5.2 丢失的数据

OHDSI 数据库中的缺失带来了微妙的挑战。本应该记录在数据库中得健康事件（例如，处方、实验室价值等）是“缺失的”。统计文献区分了诸如“完全随机缺失”、“随机缺失”和“非随机丢失”等类型的混淆，以及增加复杂性的方法去尝试解决这些类型的问题。Perkins 等人（2017）对本主题进行

了有益的介绍。

7.6 总结

在观察研究中，我们区分了三大类用例。

- 特征描述/旨在回答“他们发生了什么事？”
- 人口水平预估试图回答 “因果影响是什么”的问题
- 患者水平预测试图回答“我会发生什么？”

预测模型不是因果模型；没有理由相信对强预测因子的干预会影响结果。

有些问题无法用观察医疗健康数据来回答

7.7 练习

练习 7.1　这些问题属于哪些用例类别?

1. 计算最近接触非甾体抗炎药的患者的胃肠道出血率。

2. 根据特定患者的基线特征，计算其在下一年出现胃肠道出血的概率。

3.由于双氯芬酸与塞来昔布相比，预估胃肠道出血增加的风险。

练习 7.2　您希望评估由于双氯芬酸与非暴露（安慰剂）相比，胃肠道出血增加的风险。这可以通过观察医疗健康数据来实现吗?

建议答案见附录 E.4.

参考文献

1. Fuller, Wayne A. 2009. Measurement Error Models. Vol. 305. John Wiley & Sons.
2. Perkins, Neil J, Stephen R Cole, Ofer Harel, Eric J Tchetgen Tchetgen, BaoLuo Sun, Emily M Mitchell, and Enrique F Schisterman. 2017. “Principled Approaches to Missing Data in Epidemiologic Studies.” American Journal of Epidemiology 187 (3): 568–75.

第八章　OHDSI 分析工具

章节负责人: Martijn Schuemie & Frank DeFalco

OHDSI 提供了大量的开放源代码工具，以支持基于观察性患者数据的各种数据分析。这些工具的共同之处在于它们都可以使用“通用数据模型”（CDM）与一个或多个数据库进行交互。此外，这些工具可以使各种案例的分析流程标准化，不必从头开始，就可以通过填充标准模板来进行分析。这不仅使得数据分析更加容易，还提高了可重复性和透明度。例如，计算发病率的方法有很多种，但可以在 OHDSI 工具中选择指定的几种，并且选择相同方法将以相同的方式计算发病率。

在本章中，我们首先介绍数据分析可供选择的各种方式和策略。然后，回顾各种 OHDSI 工具以及相关用例。

8.1 分析方法

基于 CDM，我们可以使用如图 8.1 所示的三种方法对数据库进行分析研究：1）用户完全自己编写代码；2）使用 OHDSI R 包；3）交互式分析平台 ATLAS。

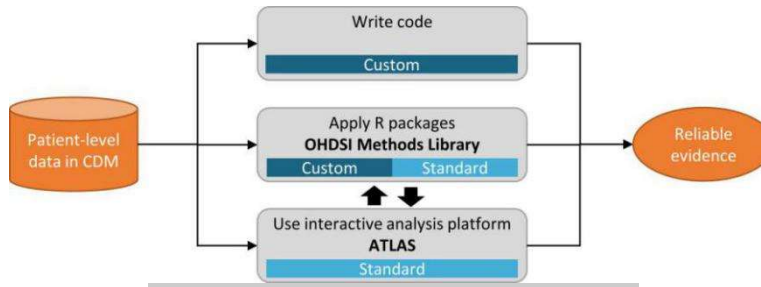


图 8.1：使用 CDM 对数据进行分析的不同方法

第一种方法不使用 OHDSI 提供的任何工具，用户用 R，SAS 或任何其他编程语言编写代码从头开始分析。这种方法提供了最大的灵活性，但需要大量的技术技能、时间和精力，并且随着分析的复杂性增加，难以避免代码错误。实际上，对于某些特定的分析场景，如果 OHDSI 没有提供适合该场景的分析工具，那么只能采用这种方法。

第二种方法使用 R 语言调用 OHDSI 方法库（即 OHDSI R 包）进行数据分析。用户可以使用 SqlRender 和 DatabaseConnector 软件包（详细介绍见第 9 章）连接各种数据库平台（例如 PostgreSQL，SQL Server 和 Oracle），也可以使用专门开发的 R 包（包含 CohortMethod 和 PatientLevelPrediction 等函数）进行基于 CDM 的高级分析。这些方法库均可以被用户自己的代码调用。这种方法仍然需要大量的专业技术知识。但与用户完全自己编写代码的方法相比，该方法因使用方法库中经过验证的内容，更高效且不易出错。

第三种方法依赖于 OHDSI 的交互式分析平台 ATLAS。ATLAS 是一个基于 Web 的图形界面工具，它让非编程人士可以高效地完成各种分析。ATLAS 利用 OHDSI 方法库，提供了一个简单的图形界面帮助用户设计数据分析流程，并且在许多情况下会生成数据分析所必要的 R 代码。但是，ATLAS 并不

支持 OHDSI 方法库中的所有方法。这意味着尽管大多数数据分析可以通过 ATLAS 来进行，但有些更加灵活的数据分析则需要采用第二种方法。

ATLAS 和 OHDSI 方法库并不是相互独立的。ATLAS 中一些比较复杂的数据分析需要通过调用 OHDSI R 包来执行。同样，OHDSI 方法库中的队列通常是在 ATLAS 中设计的。

8.2 分析策略

除了基于 CDM 的分析方法外，例如用户完全自己编写代码或使用 OHDSI R 包进行数据分析之外，还有多种使用这些分析技术生成证据的策略。图 8.2 列出了 OHDSI 中通常采用的三种策略：

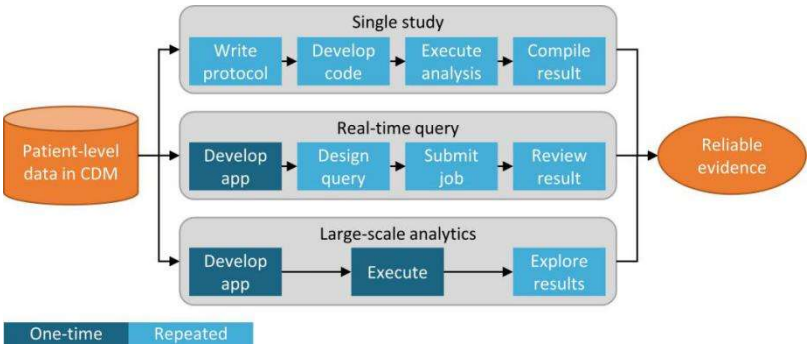


图 8.2：（临床）问题证据产生的策略

第一种策略将每个分析视为一个单独的研究。分析必须在研究方案中预先指定，以代码形式实施，然后分析数据，最后才能对结果进行解释。对于每个问题，必须重复所有步骤。这种分析的一个具体例子有“与苯妥英钠相比，左乙拉西坦引起血管性水肿的风险研究”。Duke 等人（Duke et al., 2017）首先编写了研究方案，然后使用 OHDSI 方法库开发了分析代码并在 OHDSI 的 CDM 数据上执行，然后将结果汇总在期刊上进行发表。

第二种策略是开发一种应用程序，允许用户实时或接近实时地回答特定类别的问题。开发应用程序后，用户可以交互式定义查询，提交查询并查看结果。该策略的一个具体示例是 ATLAS 中的队列定义和生成工具。该工具允许用户制定各种复杂度的队列定义，并针对数据库执行该定义，以查看有多少人符合各种纳排标准。

第三种策略同样地专注于一类问题，但随后尝试详尽地生成该类问题的所有证据。然后，用户可以根据需要通过各种交互探索证据。一个具体的例子是抑郁症治疗效果的研究（Schuemie et al., 2018b）。这项研究在四个大型观察性数据库中比较了所有抑郁症治疗的大量感兴趣的结局事件。交互式 Web 应用程序中提供了包括 177,718 个矫正经验风险比和大量的研究诊断在内的全套结果。

8.3 ATLAS

ATLAS 是 OHDSI 社区开发的一种免费的，可公开获得的基于 Web 的工具，它可以帮助设计和执行基于 CDM 格式的标准化的、患者级别的观察性数据的分析。ATLAS 与 OHDSI WebAPI 一起作为 Web 应用程序部署，通常托管在 Apache Tomcat 上。由于进行实时分析需要访问 CDM 中的患者级数据，因此通常安装在防火墙后面。但是，也有一个公共 ATLAS，尽管此 ATLAS 只能访问一些小的模

拟数据集，但仍可以达到很多测试和练习的目的。甚至完全使用 ATLAS 的公共实例来定义效果评估或预测研究，并自动生成用于执行研究的 R 代码也是可以实现的。并且，该代码可以在任何具有可用 CDM 的环境中运行，而无需安装 ATLAS 和 WebAPI。

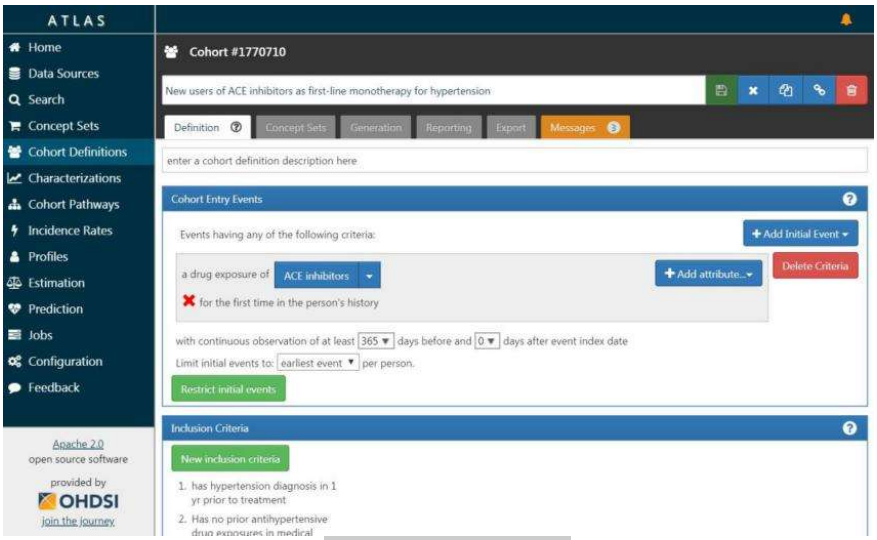


图 8.3:ATLAS 用户界面

图 8.3 提供了 ATLAS 的屏幕截图。左侧是导航栏，显示了 ATLAS 的各种功能模块：

数据源（Data Sources）

对于在 ATLAS 平台上配置的每个数据源，数据源模块提供了查看其描述性的、标准化报告的功能。此功能使用大规模分析策略：所有描述均已预先计算。数据源将在第 11 章中讨论。

词汇搜索（Vocabulary Search）

Atlas 提供了搜查和检索 OMOP 标准化术语的能力，以了解这些术语中存在哪些概念以及在针对数据源的标准化分析中如何应用这些概念。第 5 章将讨论此部分。

概念集（Concept Sets）

概念集提供了创建逻辑表达式集合的功能，这些逻辑表达式可用于标识在整个标准化分析中使用的一组概念。概念集比简单的代码或值列表更复杂。一个概念集由标准化术语中的多个概念以及逻辑指示符组成，使用者可以在词汇层级结构中指定自己感兴趣的入组或出组概念。搜索术语表，确定概念集并指定用于解析概念集的逻辑，这一点为如何定义分析计划中经常使用的晦涩医学语言提供了强大支撑。这些概念集可以保存在 ATLAS 中，然后在整个分析中用于队列定义或分析说明。

队列定义（Cohort Definitions）

队列定义是选择一组在一个持续时间内满足一个或多个条件的患者的能力。然后，这些队列可以用作所有后续分析的输入。将在第 10 章讨论。

表征（Characterizations）：表征是一种分析功能，可让您查看已定义的一个或多个队列，并总结有关这些患者人群的特征。此功能使用实时查询策略，将在第 11 章中进行讨论。

队列路径（Cohort Pathways）

队列路径模块可让您查看一个或多个人群中发生的临床事件的发生次序。此功能使用实时查询策

略，将在第 11 章中进行讨论。

发病率（Incidence Rates）

发病率模块可让您估计目标人群中结局事件的发生率。此功能使用实时查询策略，第 11 章将对此进行讨论。

个人档案（Profiles）

您可以使用 Profiles 来探索单个患者的纵向观察数据，以总结特定个体发生的事件。此功能使用实时查询策略。

人群水平的估计（Population Level Estimation）

人群水平的估计采用比较性队列设计，通过比较一个或多个目标组和对照组之间的一系列结局事件来实现群体水平的效应估计研究。此功能可以说实现了实时查询策略，因为不需要编码，将在第 12 章中进行讨论。

患者水平的预测（Patient Level Prediction）

预测功能使您可以应用机器学习算法进行患者水平的预测分析，从而可以在任何给定的目标范围内预测结局事件。由于不需要编码，因此可以说该功能实现了实时查询策略，将在第 13 章中进行讨论。

作业（Jobs）

选择“作业”菜单项以浏览通过 WebAPI 运行的进程的状态。作业通常是长期运行的流程，例如生成队列或计算特征报告。

配置（Configuration）

选择“配置”菜单项以查看已配置的数据源。

反馈（Feedback）

反馈链接会将您带到 Atlas 的问题日志，以便您可以记录新问题或搜索现有问题。如果您对新功能或优化功能有想法，请在此处为反馈给开发社区。

8.3.1 安全性

ATLAS 和 WebAPI 提供了精细的安全模型来控制对整个平台内功能或数据源的访问。该安全系统是利用 Apache Shiro 库构建的。 有关安全系统的更多信息，请参见在线 WebAPI 安全 Wiki。

8.3.2 文档

在 ATLAS GitHub 存储库 wiki 上可以找到有关 ATLAS 的文档。该 Wiki 包含有关各种应用程序功能的信息以及在线视频教程的链接。

8.3.3 如何安装

ATLAS 的安装与 OHDSI WebAPI 一起完成。在《ATLAS GitHub 存储库设置指南》和《WebAPI GitHub 存储库安装指南》中可获取每个组件的安装指南。

8.4 Methods 库

OHDSI 的 Methods 库包含了一系列开源的 R 包，如图 8.4 所示。

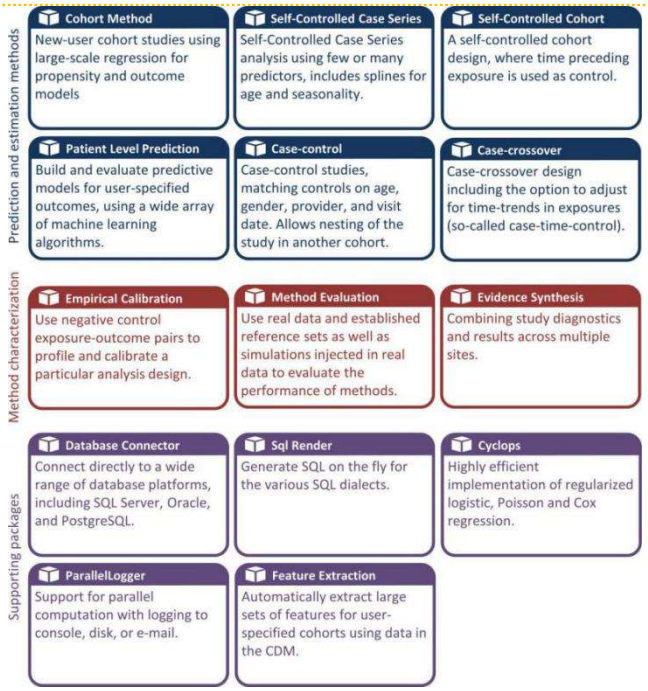


图 8.4: OHDSI 方法库中的包

这些 R 包提供了多种函数以实现完整的观察性研究，涵盖了从 CDM 数据开始，到估计、支持统计，以及制作图表等整个过程。这些 R 包可以直接对 CDM 数据操作，不仅可以为完全定制化第 9 章所描述的分析流程提供跨平台兼容性，还可以提供高级的标准化分析流程来实现人口特征分析（第 11 章）、人群水平的估计（第 12 章）和患者水平的预测（第 13 章）。Methods 库吸取以往及正在进行的研究经验为使用观察数据和观察研究设计提供了最佳的实践支持，如透明度、重复性、衡量特定背景下方法的可行性及后续的对方法所产生的评估进行经验性校正。

Methods 库已经被应用于多项发表的临床研究中（Boland et al., 2017; Duke et al., 2017; Ramcharran et al., 2017; Weinstein et al., 2017; Wang et al., 2017; Ryan et al., 2017, 2018; Vashisht et al., 2018; Yuan et al., 2018; Johnston et al., 2019），以及多项方法学研究中（Schuemie et al., 2014, 2016; Reps et al., 2018; Tian et al., 2018; Schuemie et al., 2018a,b; Reps et al., 2019）。实现 Methods 库中的所有方法的可靠性在第 17 章中描述。

8.4.1 大规模分析的支持

所有包都具有的一个特性是能够高效的运行多项分析。例如，在进行人群水平的估计的时候，CohortMethod 包允许对多个暴露和结局事件使用不同的分析设置来估计其效应值，并能够自动选择最佳方式来计算所有需要的中间及最终数据集。可以被复用的步骤将仅执行一次，例如协变量的提取、单目标组和对照组但具有多个结局事件的倾向性模型的拟合等。此外计算将尽可能的并行化以最大程度地利用计算资源。

高效的计算允许进行大规模的分析，能一次性回答多个问题，并且通过控制假设（即阴性对照）以衡量我们方法的可行性并执行经验性校正也是必不可少的（第 18 章）。

8.4.2 大数据的支持

Methods 库用于操作大数据，可以对海量的数据进行计算，这主要通过以下三条途径实现：

- 大部分的数据处理在数据库服务器上进行。单项分析通常只需要数据库中整个数据的一小部分，而通过 SqlRender 和 DatabaseConnector 包，Methods 库允许在服务器上进行高级操作来实现相关数据的预处理及提取。
- 较大的本地数据对象通过节省内存的方式进行存储。对于下载到本地机器的数据，Methods 库使用 ff 包来存储及处理大的数据对象。这使得我们可以处理比内存大得多的数据。
- 可以在有需要的时候应用于高性能计算。例如，Cyclops 包实现了一个高效的回归引擎，能够被所有 Methods 库的包调用，以实现大规模的回归计算（高维变量、大量观测）。

8.4.3 文档

R 提供了一种标准方式来为软件包提供文档。每个包都有手册（package manual），对其中的每一个函数和数据集进行说明。所有包的手册都能在 Methods 库的 GitHub 仓库中 7 在线查到。在 R 里，包的手册还可以通过问号来进行查询。例如，在载入 DatabaseConnector 包之后，输入?connect 命令就可以给出“connect”函数的文档。

除了包的手册，许多包还提供了 vignettes。vignettes 是一种长格式的文档，用于描述包是如何被用于特定的任务中。例如，这个 vignettes8 描述了如何使用 CohortMethod 包来有效的进行多种分析。Vignettes 可以在 Methods 库的 GitHub 仓库中找到，而对于 CRAN 中的包其 vignettes 可以在 CRAN 中查到。

8.4.4 系统需求

系统需求涉及两种计算环境：数据库服务器和分析工作站。

数据库服务器必须以 CDM 格式来存储观察性健康数据。Methods 库支持很多数据库管理系统，包括传统的数据库系统（(PostgreSQL，Microsoft SQL Server 和 Oracle），并行化数据库系统（Microsoft APS，IBM Netezza 和 Amazon RedShift），以及大数据平台（Hadoop through Impala 和 Google BigQuery）。

分析工作站是安装并运行 Methods 库的地方。既可以是本地机器，如某人的笔记本电脑，也可以是运行 RStudio 服务的远程服务器。不管哪种都需要安装 R，最好同时安装 RStudio。Methods 库同样需要安装 Java。分析工作站需要能够连接到数据库服务器，注意它们之间的防火墙需要开放数据库服务器的访问端口。有一些分析需要较大的计算量，因此更多的处理器核心以及扩充内存能够帮助加快分析速度。我们推荐至少要有 4 个处理器核心及 16GB 内存。

8.4.5 如何安装

下面介绍如何安装运行 OHDSI R 包所需的环境。共有 4 个软件：

1．R 是一个统计计算环境。它自带一个基本的主要是命令行的用户界面。

2．RTools 是一套在 Windows 里编译 R 包所需的程序。

3．RStudio 是使得 R 更易上手的 IDE（Integrated Development Environment）。它包括 4．代码编辑器及调试和可视化工具。请利用 RStudio 来获得更好的 R 使用体验。

Java 是运行某些 OHDSI R 包所需的计算环境。例如帮助你连接到数据库的那些包。下面我们介绍如何在 Windows 环境下安装这 4 个软件。

在 Windows 里，R 和 Java 都有 32 位和 64 位两种架构版本。如果你安装了 R 的两种架构版本，那你也必须安装 Java 的两种架构版本。推荐只安装 64 位架构版本的 R.

安装 R

打开 https://cran.r-project.org/，点击“Download R for Windows”，然后点击“base”，再点击图 8.5 所示的下载链接。

图 8.5: 从 CRAN 中下载 R

下载完成后，运行安装程序。除了以下两个地方，都使用默认安装选项：首先，最好不要安装在系统的 Program files 文件夹，而是像图 8.6 那样直接安装在 C 盘下的子文件夹中；其次，为了防止出现 R 和 Java 不同架构版本带来的问题，不要安装 32 位版本，如 Figure 8.7 所示。

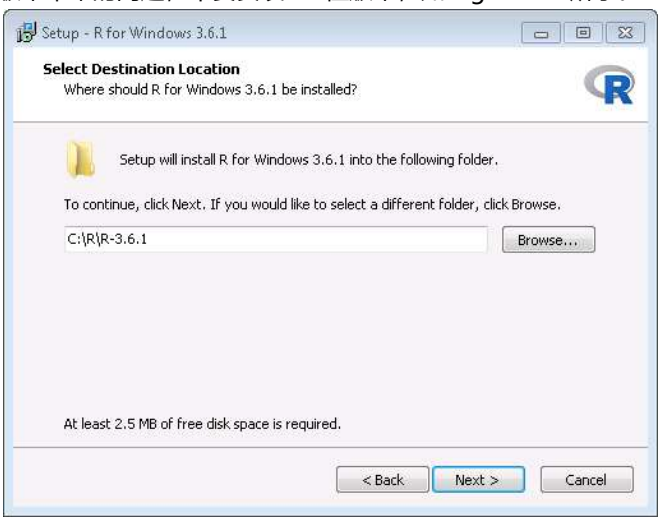


图 8.6: 设定安装 R 的路径

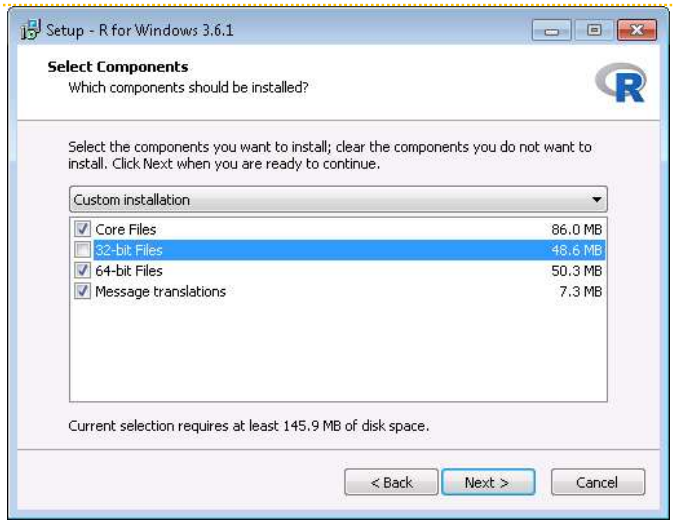


图 8.7: 禁用 32 位架构版本的 R

一旦完成后，你应当在开始菜单中能够找到 R。

安装 RTools

打开 https://cran.r-project.org/，点击 “Download R for Windows”，然后点击“Rtools”选择最新的版本下载。下载完成后，运行安装程序。请都使用默认安装选项。

安装 RStudio

打开 https://www.rstudio.com/，选择“Download RStudio”(或点击“RStudio”下方的“Download”按钮)，选择免费版本，下载 Windows 安装程序，如图 8.8 所示。

图 8.8: 下载 RStudio.

下载完成后，运行安装程序。请都使用默认安装选项。

安装 Java

打开 https://java.com/en/download/manual.jsp，选择 Windows 64 位安装程序，如图 8.9 所示。如果你安装了 32 位版本的 R，那你必须额外安装 32 位版本的 Java。

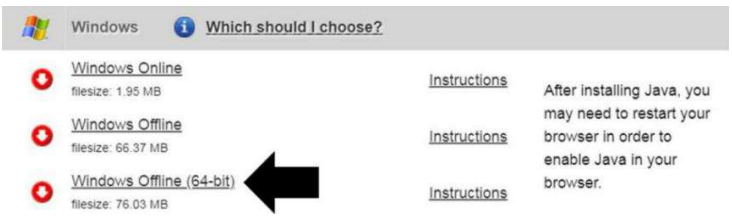


图 8.9: 下载 Java.

下载完成后，运行安装程序。

验证安装

现在你应该已准备就绪，我们来验证一下。运行 RStudio，然后输入：

```
install.packages("SqlRender")
library(SqlRender)
translate("SELECT TOP 10 * FROM person;", "postgresql")
```

8.5 部署策略

在一个组织部署整个 OHDSI 工具包，包括 ATLAS 和 Methods 库等，是一项艰巨的任务。大量的组件都需要考虑其依赖性和配置参数。因此，目前发展了两种方案来整合部署策略，使得整个工具包能够像单个包那样进行安装，这用到了 Broadsea 或 Amazon Web Services（AWS）等虚拟化方式。

8.5.1 Broadsea

Broadsea9 使用了 Docker 容器技术 10。OHDSI 工具在配置好依赖性后，被打包在了一个独立可移植的二进制 Docker 镜像文件中。这个镜像可以在 Docker 引擎中运行，生成一台安装并配置好所有软件的虚拟机。Docker 引擎在大多数的操作系统中都可以使用，包括 Microsoft Windows，MacOS 和 Linux。Broadsea 的 Docker 镜像包含了 Methods 库和 ATLAS 等主要的 OHDSI 工具。

8.5.2 Amazon AWS

Amazon 准备了两种环境可以很方便的在 AWS 云计算环境中实例化：OHDSI-in-a-Box11 和 OHDSIonAWS12。

OHDSI-in-a-Box 是专门用于学习的环境，被用于 OHDSI 社区提供的大多数教程中。它在单台便宜的 Windows 虚拟机中包含了多种 OHDSI 工具，样本数据集，RStudio 以及其它支持软件。PostgreSQL 数据库被用于存储 CDM 数据以及 ATLAS 的中间结果。OMOP CDM 数据映射以及 ETL 工具也包含在里面。OHDSI-in-a-Box 的架构如图 8.10 所示。

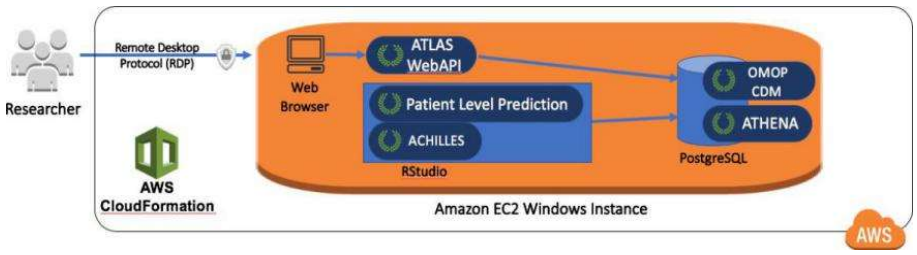


图 8.10: Amazon Web Services 中的 OHDSI-in-a-Box 架构

OHDSIonAWS 是面向企业层级、多用户、可扩展以及可容错的 OHDSI 环境，可供机构进行自己的数据分析。它包含数种样本数据集并能够自动加载机构组织自己的真实健康数据。数据存放在 Amazon Redshift 数据平台，能够被 OHDSI 工具支持。ATLAS 的中间结果存放在 PostgreSQL 数据库。在前端，用户能通过网络界面接触到 ATLAS 和 RStudio（利用 RStudio Server）。在 RStudio 里

OHDSI Methods 库已经安装好，并可以连接到数据库。OHDSIonAWS 的自动化部署是开源的，可以进行定制以包含机构组织自己的管理工具和最佳实践。OHDSIonAWS 的架构如图 8.11 所示。

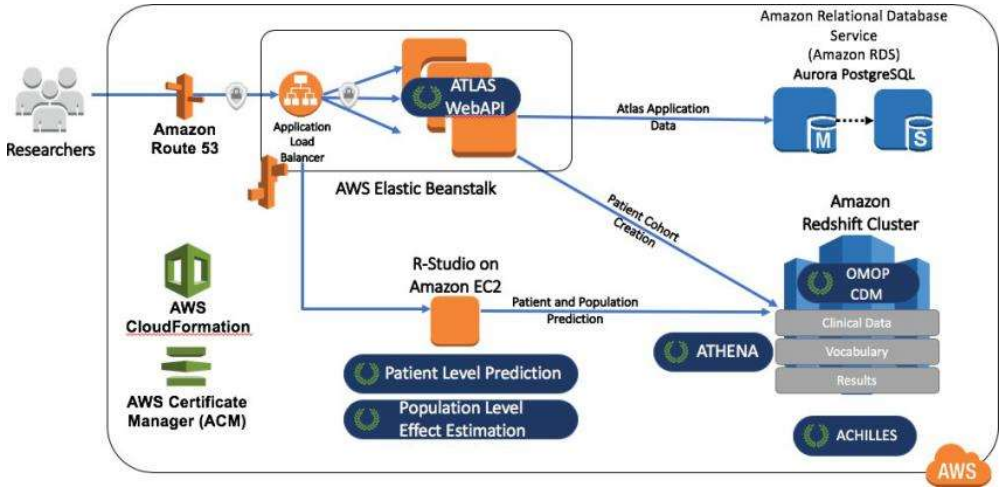


图 8.11: Amazon Web Services 中的 OHDSIonAWS 架构

8.6 总结

- 我们可以通过以下方式对 CDM 数据进行分析，包括：
  - 编写自己的代码
  - 利用 OHDSI Methods 库的 R 包编写代码
  - 使用 ATLAS 交互式分析平台
- OHDSI 工具使用不同的分析策略：
  - 单独研究
  - 实时查询
  - 大规模分析
- OHDSI 分析工具主要包含在：
  - ATLAS 交互式分析平台
  - OHDSI Methods 库的 R 包
- 可以加快 OHDSI 工具部署的几种策略。

参考文献

1. Boland, M. R., P. Parhi, L. Li, R. Miotto, R. Carroll, U. Iqbal, P. A. Nguyen, et al. 2017. “Uncovering exposures responsible for birth season - disease effects: a global study.” *J Am Med Inform Assoc*, September.
2. Duke, J. D., P. B. Ryan, M. A. Suchard, G. Hripcsak, P. Jin, C. Reich, M. S. Schwalm, et al. 2017. “Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network.” *Epilepsia* 58 (8): e101–e106.
3. Johnston, S. S., J. M. Morton, I. Kalsekar, E. M. Ammann, C. W. Hsiao, and J. Reps. 2019. “Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery.” *Value Health* 22 (5): 580–86.
4. Ramcharran, D., H. Qiu, M. J. Schuemie, and P. B. Ryan. 2017. “Atypical Antipsychotics and the Risk of Falls and Fractures Among Older Adults: An Emulation Analysis and an Evaluation of

Additional Confounding Control Strategies." *J Clin Psychopharmacol* 37 (2): 162–68.
5. Reps, J. M., P. R. Rijnbeek, and P. B. Ryan. 2019. "Identifying the DEAD: Development and Validation of a Patient-Level Model to Predict Death Status in Population-Level Claims Data." *Drug Saf*, May.
6. Reps, J. M., M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek. 2018. "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data." *Journal of the American Medical Informatics Association* 25 (8): 969–75. https://doi.org/10.1093/jamia/ocy032.
7. Ryan, P. B., J. B. Buse, M. J. Schuemie, F. DeFalco, Z. Yuan, P. E. Stang, J. A. Berlin, and N. Rosenthal. 2018. "Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: A real-world meta-analysis of 4 observational databases (OBSERVE-4D)." *Diabetes Obes Metab* 20 (11): 2585–97.
8. Ryan, P. B., M. J. Schuemie, D. Ramcharran, and P. E. Stang. 2017. "Atypical Antipsychotics and the Risks of Acute Kidney Injury and Related Outcomes Among Older Adults: A Replication Analysis and an Evaluation of Adapted Confounding Control Strategies." *Drugs Aging* 34 (3): 211–19.
9. Schuemie, M. J., G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. 2016. "Robust empirical calibration of p-values using observational data." *Stat Med* 35 (22): 3883–8.
10. Schuemie, M. 2018. "Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data." *Proc. Natl. Acad. Sci. U.S.A.* 115 (11): 2571–7.
11. Schuemie, M. J., P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan. 2014. "Interpreting observational studies: why empirical calibration is needed to correct p-values." *Stat Med* 33 (2): 209–18.
12. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. "Improving reproducibility by using high-throughput observational studies with empirical calibration." *Philos Trans A Math Phys Eng Sci* 376 (2128).
13. Tian, Y., M. J. Schuemie, and M. A. Suchard. 2018. "Evaluating large-scale propensity score performance through real-world and synthetic data experiments." *Int J Epidemiol* 47 (6): 2005–14.
14. Vashisht, R., K. Jung, A. Schuler, J. M. Banda, R. W. Park, S. Jin, L. Li, et al. 2018. "Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative." *JAMA Netw Open* 1 (4): e181755.
15. Wang, Y., M. Desai, P. B. Ryan, F. J. DeFalco, M. J. Schuemie, P. E. Stang, J. A. Berlin, and Z. Yuan. 2017. "Incidence of diabetic ketoacidosis among patients with type 2 diabetes mellitus treated with SGLT2 inhibitors and other antihyperglycemic agents." *Diabetes Res. Clin. Pract.* 128 (June): 83–90.
16. Weinstein, R. B., P. Ryan, J. A. Berlin, A. Matcho, M. Schuemie, J. Swerdel, K. Patel, and D. Fife. 2017. "Channeling in the Use of Nonprescription Paracetamol and Ibuprofen in an Electronic Medical Records Database: Evidence and Implications." *Drug Saf* 40 (12): 1279–92.
17. Yuan, Z., F. J. DeFalco, P. B. Ryan, M. J. Schuemie, P. E. Stang, J. A. Berlin, M. Desai, and N. Rosenthal. 2018. "Risk of lower extremity amputations in people with type 2 diabetes mellitus treated with sodium-glucose co-transporter-2 inhibitors in the USA: A retrospective cohort study." *Diabetes Obes Metab* 20 (3): 582–89.

39. http://data.ohdsi.org/SystematicEvidence/
40. http://www.ohdsi.org/web/atlas
41. https://github.com/OHDSI/WebAPI/wiki/Security-Configuration
42. https://github.com/OHDSI/ATLAS/wiki
43. https://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide
44. https://github.com/OHDSI/WebAPI/wiki/WebAPI-Installation-Guide
45. https://ohdsi.github.io/MethodsLibrary
46. https://ohdsi.github.io/CohortMethod/articles/MultipleAnalyses.html
47. https://github.com/OHDSI/Broadsea
48. https://www.docker.com/
49. https://github.com/OHDSI/OHDSI-in-a-Box
50. https://github.com/OHDSI/OHDSIonAWS

第九章 SQL 和 R

章节负责人: Martijn Schuemie & Peter Rijnbeek

通用数据模型（CDM）是一个关系数据库（所有数据都以字段的形式展示在数据库表格中），大多存储于 PostgreSQL, Oracle, 或者 Microsoft SQL Server 等软件平台。OHDSI 的各类工具，如 Atlas 和 Methods Library，是通过后端查询数据库来实现工具的各项功能，但我们也可在获取权限后，直接查询数据库。直接查询数据库可以帮助实现已有工具不能达成的研究需求。但是，直接查询数据库还会带来更大的出错风险，因为 OHDSI 工具通常旨在帮助指导用户进行适当的数据分析。

查询和编辑关系数据库的标准编程语言是 SQL（Structured Query Language）。虽然各类软件平台中的 SQL 基本语句都相对标准化，但各个平台的 SQL 语言也都存在少许不同。如检索 PERSON 数据表的前十行数据，可以使用下面的代码：

```
SELECT TOP 10 * FROM person;
```

然而同样的命令在 PostgreSQL 中则是：

```
SELECT * FROM person LIMIT 10;
```

在 OHDSI，与其选择使用某种特定的平台，我们希望在 OHDSI 的数据上使用统一的 SQL 语言。因此，OHDSI 开发了一个叫 SqlRender 的 R 语言软件包，可以将各个 SQL 语言翻译为本章节将讨论的任意一种语言。这个标准化后的语言，被称作 OHDSI SQL，即 SQL 语言种类中的一个子集。在本章节我们将使用 OHDSI SQL。

每个数据库平台也都有相应的软件来连接和使用 SQL 语言查询数据库。在 OHDSI，我们开发了一个叫 DatabaseConnector 的 R 语言软件包，可以连接到各个不同的数据库平台。在本章节我们也会更详细的介绍 DatabaseConnector。

所以虽然使用 OHDSI 工具来连接和查询 CDM 并非必须，但我们始终推荐使用 DatabaseConnector 和 SqlRender。这样一来，各个中心的数据库查询代码均可与其他数据中心共享，无需修改。并且，R 本身也能立即提供对数据库提取的数据进一步分析的特点，比如统计分析及生成图形。

在本章节中，读者需要对 SQL 语言具有一定的了解。首先，我们会介绍如何使用 DatabaseConnector 和 SqlRender。如果读者不想使用这两个软件包，请跳过该部分。在 9.3 中，我们将讨论如何使用 OHDSI SQL 来查询 CDM 数据。然后，我们会主要介绍如何在查询 CDM 数据时使用 OHDSI 标准术语集。其中，我们主要介绍 QueryLibrary，一个开源的对于 CDM 数据的常见查询代码集。最后，我们将使用 DatabaseConnector 和 SqlRender 来完成一个预估发生率的研究案例。

9.1 SqlRender

SqlRender 软件包可从 CRAN (the Comprehensive R Archive Network) 下载，然后用以下命令安装：

```
install.packages("SqlRender")
```

SqlRender 支持大部分平台，其中包括传统数据库系统（PostgreSQL, Microsoft SQL Server, SQLite, and Oracle），平行数据仓库（Microsoft APS, IBM Netezza, and Amazon RedShift），和大数据平台（Hadoop through Impala, and Google BigQuery）。SqlRender 的软件包自带使用指南和功能简介。接下来，我们将介绍它的主要功能。

9.1.1 SQL 参数化

该软件包其中的一项功能是为 SQL 语句提供参数化支持。通常，根据不同的参数，我们需要生成不同的 SQL 语句。SqlRender 提供了一个简单的标识语句以支持参数化。通过使用 render()功能来根据参数值生成 SQL 语句。

参数值替代

@标记了需要在 render()中被替代的参数名。在下面的例子中，变量 a 在 SQL 中被标记。当使用 render()命令时，定义了该参数的值：

```
sql <- "SELECT * FROM concept WHERE concept_id = @a;"
render(sql, a = 123)
```

```
## [1] "SELECT * FROM concept WHERE concept_id = 123;"
```

与大部分数据库管理系统提供的参数化功能不同的是，字符和表格可以跟数值一样被参数化：

```
sql <- "SELECT * FROM @x WHERE person_id = @a;"
render(sql, x = "observation", a = 123)
```

```
## [1] "SELECT * FROM observation WHERE person_id = 123;"
```

参数值可以是数字，字符，布尔型，或是被转化为以逗号分隔列表的向量：

```
sql <- "SELECT * FROM concept WHERE concept_id IN (@a);"
render(sql, a = c(123, 234, 345))
```

```
## [1] "SELECT * FROM concept WHERE concept_id IN (123,234,345);"
```

If-Then-Else

有时会根据一个或者多个变量的不同值，来决定执行或者不执行一部分的代码。这个需求可通过{Condition} ? {if true} : {if false}命令来完成。如果 Condition 的值为真或者 1，if true 部分的代码就会被使用，否则将展示 if false 部分的代码（如果存在）。

```
sql <- "SELECT * FROM cohort {@x} ? {WHERE subject_id = 1}"
render(sql, x = FALSE)
```

```
## [1] "SELECT * FROM cohort "
```

```
render(sql, x = TRUE)
```

```
## [1] "SELECT * FROM cohort WHERE subject_id = 1"
```

支持简单的比较功能：

```
sql <- "SELECT * FROM cohort {@x == 1} ? {WHERE subject_id = 1};"
render(sql, x = 1)
```

```
## [1] "SELECT * FROM cohort WHERE subject_id = 1;"
```

```
render(sql, x = 2)
```

```
## [1] "SELECT * FROM cohort ;"
```

也支持 IN 包含功能：

```
sql <- "SELECT * FROM cohort {@x IN (1,2,3)} ? {WHERE subject_id = 1};"
render(sql, x = 2)
```

```
## [1] "SELECT * FROM cohort WHERE subject_id = 1;"
```

9.1.2 翻译至其他 Sql 语言

SqlRender 的另一个功能是可将 OHDSI SQL 翻译至其他的 SQL 语言。如下：

```
sql <- "SELECT TOP 10 * FROM person;"
translate(sql, targetDialect = "postgresql")
```

```
## [1] "SELECT * FROM person LIMIT 10;"
```

targetDialect 变量可以是以下值：“oracle”，“postgresql”，“pdw”，“redshift”，“impala”，“netezza”，“bigquery”，“sqlite”，和“sql server”。

对于 SQL 恰当翻译的功能和结构上有一些局限，主要有两个原因：安装包中所具备的翻译规则有局限性；并且 some SQL features do not have an equivalent in all dialects 一些 SQL 的特性无法对等翻译到所有的方言上。这也是 OHDSI 开发了 OHDSI SQL 的原因。然而，我们会在保留 SQL 语言的基础上进行开发，避免重复造车。

尽管我们付出了各种努力，当用 OHDSI SQL 来写各平台通用的代码时，仍需注意以下几点：

翻译所支持的功能和结构

以下是通过测试可以正确翻译到各种方言的 SQL Sever 函数：

Table 9.1: 翻译所支持的功能

Function	Function	Function
ABS	EXP	RAND
ACOS	FLOOR	RANK
ASIN	GETDATE	RIGHT

ATAN	HASHBYTES*	ROUND
AVG	ISNULL	ROW_NUMBER
CAST	ISNUMERIC	RTRIM
CEILING	LEFT	SIN
CHARINDEX	LEN	SQRT
CONCAT	LOG	SQUARE
COS	LOG10	STDEV
COUNT	LOWER	SUM
COUNT_BIG	LTRIM	TAN
DATEADD	MAX	UPPER
DATEDIFF	MIN	VAR
DATEFROMPARTS	MONTH	YEAR
DATETIMEFROMPARTS	NEWID	
DAY	PI	
EOMONTH	POWER	

* 需要 Oracle 特殊权限，并不等效于 SQLite.

类似的，也可以支持多种 SQL 语言结构。以下是部分可以被正确翻译的表达形式列表：

```
-- Simple selects:
SELECT * FROM table;

-- Selects with joins:
SELECT * FROM table_1 INNER JOIN table_2 ON a = b;

-- Nested queries:
SELECT * FROM (SELECT * FROM table_1) tmp WHERE a = b;

-- Limiting to top rows:
SELECT TOP 10 * FROM table;
```

```
-- Selecting into a newtable:
SELECT * INTO new_table FROMtable;

-- Creating tables:
CREATE TABLE table (fieldINT);

-- Inserting verbatim values:
INSERT INTO other_table (field_1) VALUES (1);

-- Inserting from SELECT:
INSERT INTO other_table (field_1) SELECT value FROM table;

-- Simple drop commands:
DROP TABLE table;

-- Drop table if it exists:
IF OBJECT_ID('ACHILLES_analysis', 'U') IS NOTNULL
  DROP TABLE ACHILLES_analysis;

-- Drop temp table if it exists:
IF OBJECT_ID('tempdb..#cohorts', 'U') IS NOT NULL
  DROP TABLE #cohorts;

-- Common table expressions:
WITH cte AS (SELECT * FROM table) SELECT * FROM cte;

-- OVER clauses:
SELECT ROW_NUMBER() OVER (PARTITION BY a ORDER BY b)
  AS "Row Number" FROMtable;

-- CASE WHEN clauses:
SELECT CASE WHEN a=1 THEN a ELSE 0 END AS value FROM table;

-- UNIONs:
SELECT * FROM a UNION SELECT * FROMb;
```

字符串连接

相比于其他编程语言，字符串连接是 SQL Sever 的一个较为模糊的地方。如在 SQL Server 中，字符串连接可以是 SELECT first_name + ' ' + last_name AS full_name FROM table，但是在 PostgreSQL 和 Oracle 中则是 SELECT first_name || ' ' || last_name AS full_name FROM table。SqlRender 无法知道哪些值应该被合并。在上述例子中，由于有连接两个字符串的单引号和空格，SqlRender 可以正确翻译。但是，如果命令是 SELECT first_name + last_name AS full_name FROM table，SqlRender 就无法知道这两个值是单独的字符串，而将错误的保留代码中的加号。另一个可以代表合并字符串的命令是 VARCHAR，因此 SELECT last_name + CAST(age AS VARCHAR(3)) AS full_name FROM table 就会被正确的翻译。为避免命令模糊，建议使用 CONCAT() 命令来合并两个

或多个字符串。

数据表名和 AS 关键词

很多 SQL 语言允许使用 AS 关键词给数据表定义一个别名，但是也可选择不使用 AS。如，以下两段代码在 Server, PostgreSQL, RedShift 等语言中均可正确运行:

```
-- Using AS keyword
SELECT *
FROM my_table AS table_1
INNER JOIN (
  SELECT * FROM other_table
) AS table_2
ON table_1.person_id = table_2.person_id;

-- Not using AS keyword
SELECT *
FROM my_table table_1
INNER JOIN (
  SELECT * FROM other_table
) table_2
ON table_1.person_id = table_2.person_id;
```

然而，Oracle 会对 AS 的使用报错。在上述案例中，第一段代码即会报错。因此，在给数据表命名时，建议不适用 AS。(注意：SqlRender 并不能正确处理此情况，因为它不能分辨出何时 Oracle 可以和不可以使用 AS)

临时数据表

临时数据表可以用来存储中间步骤的结果，使用恰当则能大大提高数据查询的性能。在大部分数据平台中，临时数据表都具有许多优势：只对当前用户可见，在该工作段结束后会被自动删除，并且在用户无编辑权限下也可创建。但是，Oracle 临时数据表除了只对当前用户可见外，其他方面与永久数据表无差别。因此，在 Oracle 中，SqlRender 会通过以下两种方式建立临时数据表：

1. 在数据表名前添加随机字符，从而使得不同用户所创建的数据表名称不会重复
2. 允许用户自定义创建临时表的结构模式。

如：

```
sql <- "SELECT * FROM #children;"
translate(sql, targetDialect = "oracle", oracleTempSchema = "temp_schema")
```

```
##
[1]
"SELECT * FROM temp_schema.kefo0gk7children ;"
```

注意：用户需要 temp_schema 的编辑权。

并且由于 Oracle 限制数据表名称不能超过 30 个字符，临时数据表的名称最多可以有 22 个字符长，以留有字符空间附加 session ID。

更加需要注意的是，Oracle 中的临时数据表不会被自动删除，因此，需用户自行在完成使用后通过 TRUNCATE 和 DROP 命令删除临时表，以避免累积过多的无用数据表。

隐式转换

与其他编程语言相比，SQL Server 不太明确的几个要点之一是它允许隐式强制转换。 例如，此代码将在 SQL Server 上成功运行：

```
CREATE TABLE #temp (txt VARCHAR);
INSERT INTO #temp
SELECT '1';
SELECT * FROM #temp WHERE txt = 1;
```

即使 txt 是一个 VARCHAR 字段，并且我们将它与一个整数进行比较，SQL Server 也会自动将两者之一转换为正确的类型以进行比较。相反，其他编程语言（例如 PostgreSQL）在尝试将 VARCHAR 与 INT 进行比较时将返回错误。

因此，应该始终明确字段类型。 在上面的示例中，最后一个语句应替换为：

```
SELECT * FROM #temp WHERE txt = CAST(1 AS VARCHAR);
```

或：

```
SELECT * FROM #temp WHERE CAST(txt AS INT) = 1;
```

字符串比较中的区分大小写

一些 DBMS 平台（例如 SQL Server）始终以不区分大小写的方式执行字符串比较，而其他平台（例如 PostgreSQL）则始终区分大小写。因此，建议始终进行区分大小写的比较，并在不确定时更需要区分大小写。如，与其使用：

```
SELECT * FROM concept WHERE concep_class_id = 'Clinical Finding'
```

更推荐使用：

```
SELECT * FROM concept WHERE LOWER(concep_class_id) =
'clinical finding'
```

架构和数据库

在 SQL Server 中，表格位于架构中，而架构位于数据库中。例如，cdm_data.dbo.person 引用 cdm_data 数据库中 dbo 架构中的 person 数据表。 在其他编程语言中，即使经常存在相似的等级体系，它们的用法也有很大不同。在 SQL Server 中，每个数据库通常只有一个架构（通常称为 dbo），并且用户可以轻松地使用不同数据库中的数据。在其他平台上，例如在 PostgreSQL 中，不可能在单个时域中跨数据库使用数据，但是数据库中通常有许多架构。在 PostgreSQL 中，可以说等效于 SQL Server 的数据库就是架构。

因此，我们建议将 SQL Server 的数据库和架构连接到单个参数中，我们通常将其称为 @databaseSchema。例如，我们可以使用参数化的 SQL：

```
SELECT * FROM @databaseSchema.person
```

在 SQL Server 中，我们可以在值中包含数据库名称和架构名称：databaseSchema = “cdm_data.dbo”。在其他平台上，我们可以使用相同的代码，但是现在仅将架构指定为参数值：databaseSchema = “cdm_data”。

会报错的一种情况是 USE 命令，如 USE cdm_data.dbo; 会报错。因此，最好不要使用 USE 命令，而是指定表所在的数据库/架构。

调试参数化 SQL

调试参数化的 SQL 可能会有些复杂。只能针对数据库服务器测试呈现的 SQL，但应在参数化（预呈现）的 SQL 中更改代码。

SqlRender 程序包中包含一个 Shiny 应用程序，用于交互式编辑源 SQL 并生成渲染的和转换的 SQL。 该应用可以使用以下命令启动：

```
launchSqlRenderDeveloper()
```

这个命令将会打开默认浏览器，如图 9.1 所示。该应用程序也在网络上公开可用。[51]

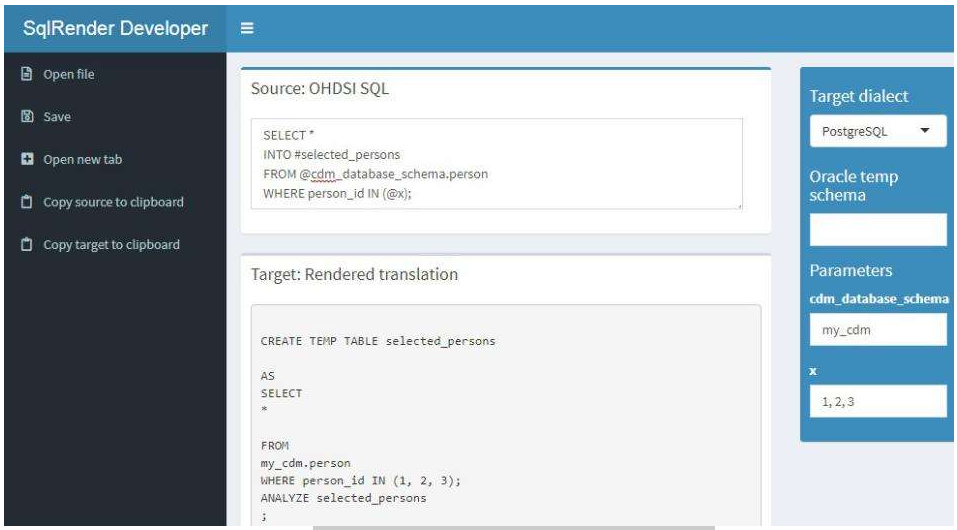


图 9.1: SqlDeveloper Shiny 应用程序

在应用程序中，您可以输入 OHDSI SQL，选择目标编程语言以及为 SQL 中显示的参数提供值，翻译将自动显示在底部。

9.2 DatabaseConnector

DatabaseConnector 是一个 R 软件包，用于使用 Java 的 JDBC 驱动程序连接到各种数据库平台。DatabaseConnector 软件包开放在 CRAN（the Comprehensive R Archive Network），可以使用以下方式安装：

```
install.packages("DatabaseConnector")
```

DatabaseConnector 支持广泛的技术平台，包括传统的数据库系统（PostgreSQL，Microsoft SQL Server，SQLite 和 Oracle），并行数据仓库（Microsoft APS，IBM Netezza 和 Amazon RedShift）以及大数据平台（通过 Impala 的 Hadoop 和 Google BigQuery）。该软件包已经包含了大多数驱动程序，但是由于许可原因，BigQuery，Netezza 和 Impala 的驱动程序不包括在内，必须由用户去获得。键入？ jdbcDrivers 可以获取有关如何下载这些驱动程序的说明。下载后，可以使用 connect，

dbConnect 和 createConnectionDetails 函数的 pathToDriver 参数。

9.2.1 创建连接

要连接到数据库，我们指定许多详细信息，例如数据库平台，服务器的位置，用户名和密码。我们可以调用 connect 函数并直接指定这些详细信息：

```
conn <- connect(dbms ="postgresql",
                server = "localhost/postgres",
                user = "joe",
                password ="secret",
                schema = "cdm")
```

```
## Connecting using PostgreSQL driver
```

有关每个平台需要哪些详细信息，请使用？connect 命令来获取。并且 请在使用完毕后断开连接：

```
disconnect(conn)
```

注意，如果更方便，也可以提供 JDBC 连接字符串，而不是提供服务器名称：

```
connString <- "jdbc:postgresql://localhost:5432/postgres"

conn <- connect(dbms = "postgresql",
                connectionString = connString,
                user = "joe",
```

```
## Connecting using PostgreSQL driver
```

有时我们可能想先指定连接的详细信息，之后再进行连接。例如，当在函数内部建立连接并且需要将详细信息作为参数传递时，这可能会很方便。为此，我们可以使用 createConnectionDetails 函数：

```
## Connecting using PostgreSQL driver
```

```
details <- createConnectionDetails(dbms = "postgresql",
                                   server = "localhost/postgres",
                                   user = "joe",
                                   password ="secret",
                                   schema = "cdm")
conn <- connect(details)
```

9.2.2 查询

查询数据库的主要使用 querySql 和 executeSql 函数。 这些函数之间的区别在于 querySql 的数据由数据库返回，并且一次只能处理一个 SQL 语句。相反，executeSql 并不返回数据，而是在单个 SQL 字符串中接受多个 SQL 语句。

如下：

```
querySql(conn, "SELECT TOP 3 * FROM person")
```

```
## person_id gender_concept_id year_of_birth
## 1         1                 8507           1975
## 2         2                 8507           1976
## 3         3                 8507           1977
```

```
executeSql(conn, "TRUNCATE TABLE foo; DROP TABLE foo;")
```

两种功能均提供了详细的错误报告：服务器报出错误时，错误消息和有问题的 SQL 片段将被写入文本文件中，以实现更好的调试。 默认情况下，executeSql 函数还会默认显示一个进度条，指示已执行的 SQL 语句的百分比。如果不需要这些属性，该软件包还提供了 lowLevelQuerySql 和 lowLevelExecuteSql 函数。

9.2.3 使用 Ffdf 对象查询

有时，要从数据库中获取的数据太大而无法放入内存。如第 8.4.2 节所述，在这种情况下，我们可以使用 ff 包将 R 数据对象存储在文件中，并像在内存中可用一样使用它们。DatabaseConnector 可以将数据直接下载到 ffdf 对象中：

```
x <- querySql.ffdf(conn, "SELECT * FROM person")
```

X 就成为了 ffdf 对象。

9.2.4 使用相同的 SQL 查询不同的平台

可以使用以下便利的功能，这些功能首先调用 SqlRender 包中的 render 和 translate 函数：renderTranslateExecuteSql，renderTranslateQuerySql，renderTranslateQuerySql.ffdf。例如：

```
x <- renderTranslateQuerySql(conn,
                             sql = "SELECT TOP 10 * FROM @schema.person",
                             schema = "cdm_synpuf")
```

请注意，在 PostgreSQL 上，特定于 SQL Server 的“TOP 10”语法将转换为例如“LIMIT 10”，并且 SQL 参数@schema 会使用“cdm_synpuf”作为值。

9.2.5 插入表格

尽管也可以通过使用 executeSql 函数发送 SQL 语句在数据库中插入数据，但是使用 insertTable 函数通常更方便，更快捷（由于某些优化）：

```
data(mtcars)
insertTable(conn, "mtcars", mtcars, createTable = TRUE)
```

在示例中，我们将数据框 mtcars 上传到服务器上名为“ mtcars”的表中，该表将自动创建。

9.3 查询 CDM

在以下示例中，我们使用 OHDSI SQL 查询符合 CDM 的数据库。 这些查询使用@cdm 表示存储 CDM 的数据库架构。

我们可以从查询数据库中的人数开始：

```
SELECT COUNT(*) AS person_count FROM
@cdm.person;
```

PERSON_COUNT
26299001

也许我们对观察期的平均长度感兴趣：

```
SELECT AVG(DATEDIFF(DAY,
                    observation_period_start_date,
                    observation_period_end_date) / 365.25) AS
num_years
FROM @cdm.observation_period;
```

NUM_YEARS
1.980803

我们可以联接表以产生其他统计信息。通常要求联接来自多个表中的特定字段的值相等。例如，在这里，我们将使用 PERSON_ID 字段来连接 PERSON 表和 OBSERVATION_PERIOD 表。换句话说，联接的结果是一个新的类似表的集合，该集合具有两个表的所有字段，但是在所有行中，两个表的 PERSON_ID 字段必须具有相同的值。例如，我们现在可以通过使用 OBSERVATION_PERIOD 表中的 OBSERVATION_PERIOD_END_DATE 字段以及 PERSON 表的

```
SELECT MAX(YEAR(observation_period_end_date) -
           year_of_birth) AS max_age
FROM @cdm.person
INNER JOIN @cdm.observation_period
  ON person.person_id =observation_period.person_id;
```

year_of_birth 字段来计算观察期结束时的最大年龄：

MAX_AGE
90

如果需要确定观察开始时的年龄分布，需要一个更复杂的查询。在此查询中，我们首先将 PERSON 表和 OBSERVATION_PERIOD 表联接在一起，来计算在观察开始时的年龄。我们算出此联接集中按年龄排序的序号，并将其存储为 order_nr。 由于我们想多次使用此联接表的结果，所以我们将其定义为我们称为“年龄”的通用表表达式（CTE）（使用 WITH ... AS 定义），这意味着我们可以将年龄称为现

```
WITH ages
AS (
    SELECT age,
        ROW_NUMBER() OVER (
            ORDER BY age
            ) order_nr
    FROM (
        SELECT YEAR(observation_period_start_date) - year_of_birth AS age
        FROM @cdm.person
        INNER JOIN @cdm.observation_period
            ON person.person_id = observation_period.person_id
        ) age_computed
    )
```

有数据表。我们计算年龄的行数以产生“ n”，然后对于每个分位数区间，找到 order_nr <.50 * n。最小和最大年龄分别以以下方式计算：

MIN_AGE	Q25_AGE	MEDIAN_AGE	Q75_AGE	MAX_AGE
0	6	17	34	90

也可以在 R 中执行更复杂的计算。例如，我们可以使用以下 R 代码获得相同的答案：

```
sql <- "SELECT YEAR(observation_period_start_date) -
              year_of_birth AS age
FROM @cdm.person
INNER JOIN @cdm.observation_period
  ON person.person_id = observation_period.person_id;"
age <- renderTranslateQuerySql(conn, sql, cdm = "cdm")
quantile(age[, 1], c(0, 0.25, 0.5, 0.75, 1))
```

```
##   0%  25%  50%  75% 100%
##    0    6   17   34   90
```

在这里，我们计算服务器上的年龄，下载所有年龄，然后计算年龄分布。 但是，这需要从数据库服务器下载数百万行数据，因此效率不是很高。您将需要根据具体情况决定是在 SQL 还是 R 中执行计算。

可查询在 CDM 中的原始数据值。例如我们可以用以下方法检索前 10 个最常见的疾病源代码：

```
SELECT TOP 10 condition_source_value,
  COUNT(*) AS code_count
FROM  @cdm.condition_occurrence
GROUP BY condition_source_value
ORDER BY -COUNT(*);
```

CONDITION_SOURCE_VALUE	CODE_COUNT
4019	49094668
25000	36149139
78099	28908399
319	25798284
31401	22547122
317	22453999
311	19626574
496	19570098
I10	19453451
3180	18973883

在这里，我们根据 CONDITION_SOURCE_VALUE 字段的值将 CONDITION_OCCURRENCE 表中的记录分组，并计算每组中的记录数。我们检索 CONDITION_SOURCE_VALUE 和计数，然后按计数对它进行反向排序。

9.4 使用词汇进行查询

许多操作需要使用词汇表。词汇表是 CDM 的一部分，因此可以通过 SQL 查询使用。在这里，我们展示了如何将词汇（Vocabulary）的查询与 CDM 的查询结合起来。CDM 表中的许多列都是 CONCEPT ID。 例如，我们想要按性别分类计算数据库中的人数，那么用 GENDER_CONCEPT_ID 找到性别的概念名称会很方便：

```
SELECT COUNT(*) AS
  subject_count,
  concept_name
FROM @cdm.person
INNER JOIN @cdm.concept
  ON person.gender_concept_id =concept.concept_id
GROUP BY concept_name;
```

SUBJECT_COUNT	CONCEPT_NAME
14927548	FEMALE
11371453	MALE

词汇表的一个非常强大的功能是其层次结构。 一个非常常见的查询是寻找一个特定概念及其所有子概念。 例如，假如我们想要计算包含布洛芬成分的处方数量：

```
SELECT COUNT(*) AS prescription_count
FROM @cdm.drug_exposure
INNER JOIN @cdm.concept_ancestor
  ON drug_concept_id =descendant_concept_id
INNER JOIN @cdm.concept ingredient
 ON ancestor_concept_id = ingredient.concept_id
WHERE LOWER(ingredient.concept_name)='ibuprofen'
 AND ingredient.concept_class_id = 'Ingredient'
AND ingredient.standard_concept = 'S';
```

PRESCRIPTION_COUNT
26871214

9.5 QueryLibrary

QueryLibrary 是 CDM 的常用 SQL 查询库。它可以作为在线应用程序[52]使用，如图 9.2 所示，也可以作为 R 软件包使用[53]。

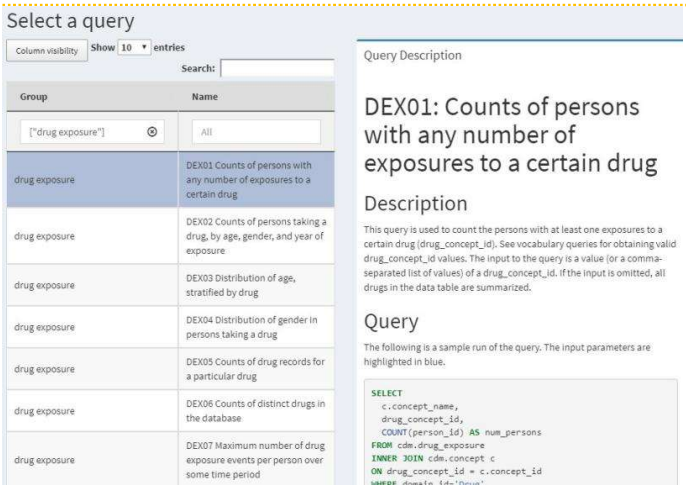


Figure 9.2: QueryLibrary: 针对 CDM 的 SQL 查询库.

该库的目的是帮助新用户学习如何查询 CDM。该库中的查询已由 OHDSI 协会审查和批准。查询库主要用于培训目的，但它对于有经验的用户也是宝贵的资源。

QueryLibrary 利用 SqlRender 输出用户所选择的 SQL 语言。用户还可以指定 CDM 数据库架构，词汇数据库架构（如果单独）和 Oracle 临时数据库架构（如果需要），因此查询将通过这些设置自动呈现。

9.6 设计一个简单的研究

9.6.1 问题定义

血管性水肿是 ACE 抑制剂（ACEi）的众所周知的副作用。 Slater et al.（1988）估计 ACEi 治疗第一周的血管性水肿发病率为每周每 3,000 名患者中就有 1 例。在这里，我们试图复制这一发现，并按年龄段和性别进行分层分析。为简单起见，我们专注于一种 ACEi：（Lisinopril）赖诺普利。我们要回答的问题是:

赖诺普利治疗开始后第一周,按年龄段和性别分类血管性水肿发病率是多少?

9.6.2 用药

我们定义用药为第一次使用赖诺普利。所谓第一次,我们是指在此之前没有使用过赖诺普利。首次使用之前，我们需要 365 天的连续观察时间。

9.6.3 结果

我们将血管性水肿定义为,在住院或急诊期间出现的任何血管性水肿的诊断代码。

9.6.4 风险时段

一旦治疗开始，我们就计算治疗开始后的第一周的发病率，而不用考虑患者是否整周都在用药。

9.7 使用 SQL 和 R 实施研究

尽管并没有要求我们一定要遵守 OHDSI 工具的使用惯例，但遵循相同的原则将很有帮助。这里，我们将使用 SQL 填充 COHORT 表，类似于 OHDSI 工具的工作方式。CDM 中有 COHORT 表的定义，我们将使用其中的一部分预定义。首先，我们必须在具有编辑权限的数据库架构中创建 COHORT 表，这可能与 CDM 保存数据的数据库架构不同。

```
library(DatabaseConnector)
conn <- connect(dbms = "postgresql",
                server = "localhost/postgres",
                user = "joe",
                password = "secret")

cdmDbSchema <- "cdm"
cohortDbSchema <- "scratch"
cohortTable <- "my_cohorts"
```

```
sql <- "
CREATE TABLE @cohort_db_schema.@cohort_table
  ( cohort_definition_id INT,
  cohort_start_date DATE,
  cohort_end_date DATE,
  subject_id BIGINT
)
;
"
renderTranslateExecuteSql(conn, sql,
                          cohort_db_schema = cohortDbSchema,
                          cohort_table = cohortTable)
```

在这里，我们已经对数据库架构和表名进行了参数化，因此我们可以轻松地将它们用在不同的环境。以上语句的结果是会在数据库服务器上建立一个空表。

9.7.1 用药队列

接下来，我们创建用药队列，并将其放入到 COHORT 表中：

```
sql <- "
INSERT INTO @cohort_db_schema.@cohort_table
  ( cohort_definition_id,
  cohort_start_date,
  cohort_end_date,
  subject_id
)
SELECT 1 AS cohort_definition_id,
  cohort_start_date,
  cohort_end_date,
  subject_id
FROM (
  SELECT drug_era_start_date AS cohort_start_date,
    drug_era_end_date AS cohort_end_date, person_id
    AS subject_id
  FROM (
    SELECT drug_era_start_date,
      drug_era_end_date, person_id,
      ROW_NUMBER() OVER (
        PARTITION BY person_id
            ORDER BY drug_era_start_date
      ) order_nr
    FROM @cdm_db_schema.drug_era

    WHERE drug_concept_id = 1308216 --Lisinopril
  ) ordered_exposures
  WHERE order_nr = 1
```

我们使用 DRUG_ERA 表，这是自动从 DRUG-EXPOSURE 表自动派生出的 CDM 中的标准表。DRUG_ERA 表包含了暴露在连续用药暴露的时间。这样我们搜索 Lisinopril，就会自动识别所有还有 Lisinopril 的暴露。然后我们将每个人的首次用药时间与 OBSERVATION_PERIOD 表结合起来，由于一个人可以有多个观察期，因此必须确保我们仅关联包含用药时段的观察期。然后，我们需要在 OBSERVATION_PERIOD_START_DATE 和 COHORT_START_DATE 之间至少有 365 天。

9.7.2 结果队列

最后，我们来建立结果队列：

```
sql <- "
INSERT INTO @cohort_db_schema.@cohort_table
 ( cohort_definition_id,
 cohort_start_date,
cohort_end_date, subject_id
)
SELECT 2 AS cohort_definition_id,
  cohort_start_date,
  cohort_end_date,
  subject_id
FROM (
  SELECT DISTINCT person_id AS subject_id,
    condition_start_date AS cohort_start_date,
    condition_end_date AS cohort_end_date
  FROM @cdm_db_schema.condition_occurrence
  INNER JOIN @cdm_db_schema.concept_ancestor
    ON condition_concept_id = descendant_concept_id
  WHERE ancestor_concept_id = 432791 -- Angioedema
) distinct_occurrence
```

```
INNER JOIN @cdm_db_schema.visit_occurrence ON
  subject_id =person_id
  AND visit_start_date <= cohort_start_date AND
  visit_end_date >= cohort_start_date
WHERE visit_concept_id IN (262, 9203,
    9201) -- Inpatient or ER;
"

renderTranslateExecuteSql(conn, sql,
                          cohort_db_schema = cohortDbSchema,
                          cohort_table = cohortTable,
                          cdm_db_schema =cdmDbSchema)
```

将 CONDITION_OCCURRENCE 表连接 CONCEPT_ANCESTOR 表，找出所有血管性水肿及其子概念发生的记录。由于同一天血管性水肿被多次诊断的可能性很高，而并非真正发生了多次血管性水肿事件，我们使用 SQL 中的 DISTINCT 来确保每天只选择一次记录。 然后我们将这些记录与 VISIT_OCCURRENCE 表联接，以确保诊断是在住院或急诊期间做出的。

9.7.3 发病率计算

现在我们的队列做好了，我们可以计算出按年龄段和性别分层的发病率：

```
sql <- " WITH
tar AS (
  SELECT concept_name AS gender,
    FLOOR((YEAR(cohort_start_date)  -
          year_of_birth) / 10) AS age,
    subject_id,
    cohort_start_date,
    CASE WHEN DATEADD(DAY, 7, cohort_start_date) >
      observation_period_end_date
    THEN observation_period_end_date
    ELSE DATEADD(DAY, 7, cohort_start_date) END
    AS cohort_end_date
  FROM @cohort_db_schema.@cohort_table
  INNER JOIN @cdm_db_schema.observation_period ON
    subject_id = observation_period.person_id
      AND observation_period_start_date < cohort_start_date
      AND observation_period_end_date > cohort_start_date
   INNER JOIN @cdm_db_schema.person
       ON subject_id = person.person_id
   INNER JOIN@cdm_db_schema.concept
       ON gender_concept_id = concept_id
   WHERE cohort_definition_id = 1 -- Exposure
)

       ) events

results <- renderTranslateQuerySql(conn, sql,
                                   cohort_db_schema = cohortDbSchema,
                                   cohort_table = cohortTable,
                                   cdm_db_schema = cdmDbSchema,
                                   snakeCaseToCamelCase = TRUE)
```

我们首先创建“tar”， 即包含所有具有适当风险时段的事件的一种 CTE。请注意，我们在 observation_periods _end_date 中截断风险时间。我们按性别和 10 年的时间段进行计算。使用 CTE 的优点是我们可以多次使用同一查询的中间结果。这个案例中，我们使用它来计算风险时段的总数以及在风险时段内发生血管性水肿事件的数量。

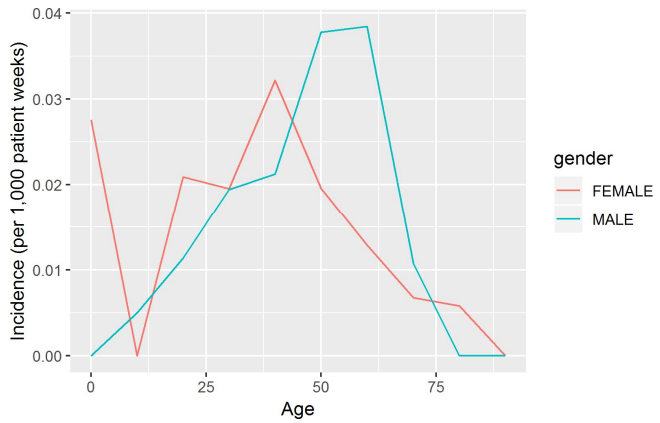
我们使用 snakeCaseToCamelCase = TRUE，因为在 SQL 中，我们倾向于用 s nake_case 作为表列的名称(因为 SQL 不区分大小写)，而在 R 中，我们倾向于使用 camelCase（因为 R 区分大小写)。而 results 数据表列名称现在将在 camelCase 显示。

借助 ggplot2 软件包，我们可以轻松地绘制结果：

```
# Compute incidence rate (IR):
results$ir <- 1000 * results$events / results$days / 7

# Fix age scale:
results$age <- results$age * 10

library(ggplot2)
ggplot(results, aes(x = age, y = ir, group = gender, color = gender)) +
  geom_line() +
  xlab("Age") +
  ylab("Incidence (per 1,000 patient weeks)")
```



9.7.4 清理

不要忘了清理我们创建的表，并关闭连接：

```
sql <- "
TRUNCATE TABLE @cohort_db_schema.@cohort_table;
DROP TABLE @cohort_db_schema.@cohort_table;
"
renderTranslateExecuteSql(conn, sql,
                          cohort_db_schema = cohortDbSchema,
                          cohort_table = cohortTable)
disconnect(conn)
```

9.7.5 相容性

由于我们将 OHDSI SQL 与 DatabaseConnector 和 SqlRender 一起使用，因此我们在本章中用的代码适用于任何 OHDSI 支持的数据库平台。

请注意，出于演示目的，我们选择使用手工 SQL 创建队列。在 ATLAS 中构建队列定义，并使用 ATLAS 产生的 SQL 来建立队列，可能会更加方便。 ATLAS 还同时会生成 OHDSI SQL，因此可以轻松地与 SqlRender 和 DatabaseConnector 一起使用。

9.8 总结

– SQL（结构化查询语言）是用于查询符合通用数据模型（CDM）的数据库的标准语言。
– 不同的数据库平台具有不同的 SQL 语言，并且需要不同的工具来查询它们。
– SqlRender 和 DatabaseConnector R 软件包提供了一种统一的方式来查询 CDM 中的数据，从而使同一个分析代码无需修改即可在不同的环境中运行。
– 一起使用 R 和 SQL，我们可以实现 OHDSI 工具不支持的自定义分析。
– QueryLibrary 为 CDM 提供了一组可重复使用的 SQL 查询。

9.9 练习

预备知识

对于这些练习，我们假设您已按照第 8.4.5 节中的说明安装了 R，R-Studio 和 Java。 还需要 SqlRender，DatabaseConnector 和 Eunomia 软件包，这些软件包可以使用以下方法安装:

```
install.packages(c("SqlRender", "DatabaseConnector", "devtools"))
devtools::install_github("ohdsi/Eunomia",ref="v1.0.0")
```

Eunomia 软件包在 CDM 中提供了模拟的数据集，该数据集将在本地 R session 中运行。 可以使用以下方法获取连接详细信息:

```
connectionDetails <-Eunomia::getEunomiaConnectionDetails()
```

CDM 数据库架构为“main”。

练习 9.1 使用 SQL 和 R 计算数据库中的人数。

练习 9.2 使用 SQL 和 R，计算至少有一次 celecoxib 处方的人数。

练习 9.3 使用 SQL 和 R，计算用 celecoxib 期间有多少次胃肠道出血的诊断。(提示：胃肠道出血的概念 ID 为 192671。)

参考答案可以在附录 E.5 中找到。

参考文献

1. Slater, Eve E., Debora D. Merrill, Harry A. Guess, Peter J. Roylance, Warren D. Cooper, William H. W. Inman, and Pamela W. Ewan. 1988. “Clinical Profile of Angioedema Associated With Angiotensin Converting-Enzyme Inhibition.” *JAMA* 260 (7): 967–70. https://doi.org/10.1001/jama.1988.03410070095035.

51. http://data.ohdsi.org/SqlDeveloper/
52. http://data.ohdsi.org/QueryLibrary
53. https://github.com/OHDSI/QueryLibrary

第十章 队列定义

章节负责人: Kristin Kotska

健康观察型数据，也称为真实世界数据，是与患者健康状况相关或与患者医疗服务相关的各种来源的数据。OHDSI 数据管理员（将他们的数据转化为通用数据模型的 OHDSI 合作者）可以从许多来源获得数据，包括电子健康记录（EHR），和收费健康保险理赔和收费数据，产品和疾病注册数据库，患者生成的数据等。患者生成的数据包括家用设备中的数据，以及其他记录健康状况的数据，如移动设备等。由于这些数据并非出于研究目的而收集，因此这些数据可能并不能直接显示我们感兴趣的临床数据要素。

例如，健康保险理赔数据库旨在获取针对于某种症状（如血管性水肿）提供的所有医疗服务，并且仅针对此目的获取实际状况信息，用于合理报销相关费用。如果我们希望将此类观测数据用于研究目的，我们通常不得不编写一些算法来获得数据中我们真正感兴趣的内容。换句话说，我们通常需要运用某些概念去表现临床事件从而去建立队列。因此，如果我们想在保险理赔数据库中识别血管性水肿案例，我们可以把急诊室记录中的血管性水肿的诊断代码作为一种逻辑定义，以区别于随访照护记录中仅描述过去发生的血管性水肿事件。类似的做法也适用于 EHR 中记录的常规医疗交互过程中获得的数据。此外，当数据被用于其他目的时，我们必须了解每个数据库设计时的初衷。每次设计一个研究时，我们必须仔细考虑队列在各种医疗数据中可能的细微差异。

本章解释旨在解释创建和分享队列定义的含义，建立队列的方法，以及如何用 ATLAS 或 SQL 建立你自己的队列。

10.1 队列是什么

在 OHDSI 研究中，我们将队列定义为在一段时间内满足一个或者多个入选标准的一组人。术语“队列”与术语“表型”相似。队列在整个 OHDSI 分析工具和网络研究中都是研究问题的主要构建模块。例如，在一项使用 ACEI 的人群中预测血管性水肿风险的研究中，我们定义了两个队列：结果队列（血管性水肿）和目标队列（使用 ACEI 的人群）。OHDSI 中队列定义的一个重要方面是队列通常与研究中的其他队列独立定义，因此可以重复使用。例如，在我们的例子中，血管性水肿队列定义了人群中所有的血管性水肿事件，并且包括目标人群以外的事件。在分析需要的时候，分析工具会采用这两个队列的交集进行分析。这样做的好处是我们也可以在不同的分析中使用相同的血管性水肿队列定义，例如，将 ACEI 与其他暴露进行比较的评估研究。队列定义因研究而异，具体取决于感兴趣的研究问题。

队列是在一段时间内满足一个或多个入选标准的一组人。

重要的一点是，在 OHDSI 中使用的队列可能与本领域其他人使用的队列定义不同。例如，在很多同行评议的科研文章中，队列被定义为类似于特定临床代码的编码集（如 ICD-9/ICD-10， NDC，HCPCS 等）。虽然编码集是构建队列时的重要组成部分，但队列的定义不是由编码集定义的。队列需要有使用标准编码集的特定逻辑（如，是首次出现 ICD-9/ICD-10 代码吗？还是任何一次?）。一个定义

明确的队列定义了纳入患者如何纳入队列以及患者如何退出队列。

用 OHDSI 的定义队列有一些细微差别，包括：

- 一个人可能属于多个队列
- 一个人可能在多个不同的时间段属于同一队列
- 同一时间段内，一个人可能不在同一队列里
- 一个队列可能有 0 个或者多个成员

建立队列的方法主要有两种：

1．**基于规则的队列定义：**使用明确的规则来描述在队列中的患者。定义这些规则在很大程度上取决于设计队列的个人的专业领域知识以及他们对目标治疗领域的知识来建立队列纳入标准。

2．**基于概率的队列定义：**使用概率模型来计算患者在队列中的可能性。可以设定某个阈值将该概率转化为是否分类，或者在某些研究设计中可以按照概率使用。通常使用机器学习模型（如逻辑回归）在一些样本数据上训练概率模型，来自动识别相关患者的特征。

下一节将更详细地讨论这些方法。

10.2 基于规则的队列定义

基于规则的队列定义是在特定时间段内（如“在最近六个月出现这种症状”）明确规定一个或者多个入选标准（如“血管性水肿患者”）。

用来规定这些标准的标准组成部分是：

- 域：存储数据的通用数据模型的域（如“操作”，“药物暴露”）定义了临床信息的类型并且在该通用数据模型表中展示的概念。在章节 4.2.4 中会更详细地讨论域。
- 概念集：一种与数据无关的表达方式，定义了一个或者多个所关注的临床特征的标准概念。因为这些概念集代表了临床特征在医学概念集中的映射，概念集可跨不同的观察健康数据使用。概念集将在章节 10.3 中深入讨论。
- 特定于域的属性：与临床特征相关的其他属性（如，药物暴露域中的“DAYS_SUPPLY”，检验域中的“VALUE_AS_NUMBER”或者“RANGE_HIGH”）。
- 序纳入标准时序逻辑：评估纳入标准与事件之间关系的时间间隔（如：所述的症状需要在暴露开始之前 365 天发生）。

在建立队列定义时，你可能会发现将域比作构建队列属性的构件模块会有所帮助（见图 10.1）如果你对每个方面中所存储的内容感到困惑，可以参考第四章的通用数据模型。

当创建队列定义时，你需要问自己以下问题：

- 哪些初始事件定义了队列入组的时间?
- 初始事件采用了哪些纳入标准?
- 什么定义了队列排除的时间?

队列入组事件：队列入组事件（初始事件）定义了人群进入队列的时间，称为队列的索引日期。队列入组事件可以是 CDM 中记录的任何事件，如药物暴露、症状、检验、访视等。初始事件由存储数据的 CDM 域定义（如 PROCEDURE_OCCURRENCE，DRUG_EXPOSURE 等），构建用于识别临床活动术语集的概念集（如针对于症状的 SNOMED 术语集，针对药物的 RxNorm 术语集），以及其他特定属性（例如发生年龄，首次诊断/手术等，指定的开始和结束日期，指定的访问类型或标准、天数等）。发

生队列入组事件的人群称为初始事件队列。

纳入标准：纳入标准是应用于初始事件队列上的用于进一步限定人群的条件。每个纳入标准是由存储数据的 CDM 域，代表临床活动术语集的概念集，特定于域的属性（如天数，就诊类型等），以及相对于队列索引日期的时间逻辑来定义的。纳入标准可以评估每个标准对于初始事件队列中人员流失的影响。符合条件的队列定义为初始事件队列中满足纳入标准的所有人群。

排除标准：队列退出事件表示一个人不再有队列资格。队列退出可以通过不同的方式来定义，例如观察期结束，相对于初始事件的固定持续时间，一系列相关观测中的最后一个事件（例如持续的药物暴露）或者通过观察审查排除标准期的审查。队列排除标准影响一个人在不同的时间间隔内是否可以多次属于同一个队列。

图 10.1 ： 队列定义的构件模块

在 OHDSI 工具中，纳入标准和排除标准之间没有区别。所有的条件均被定为纳入标准。例如，可以将排除标准“排除先前患有高血压的患者”定义为纳入标准“包括 0 例先前患有高血压的患者”。

10.3 概念集

概念集是一系列在分析中可被重复使用的概念。可以将其视为观测性研究中的经常使用的标准化的、计算机可执行的编码列表。概念集包括列表和以下属性：

- 排除：术语从概念集中排除此概念（及其任何后代）。
- 子集：需要同时考虑这个概念和它的所有的子集。
- 映射：允许搜索非标准的概念。

例如，概念集可以包括两个概念，如表 10.1 所示。这里，我们包括概念 4329847 （“心肌梗塞”）及其所有子集，但不包括概念 314666（“旧心肌梗塞”）及其所有子集.

表 10.1: 示例概念集

Concept Id	Concept Name	Excluded	Descendants	Mapped
4329847	Myocardial infarction	NO	YES	NO
314666	Old myocardial infarction	YES	YES	NO

如图 10.2 所示，这个概念集包括“心肌梗塞”及其所有子集，但“旧心肌梗塞”及其所有子集除外。此概念集包括总共将近一百个标准概念。这些标准概念又反映了可能出现在各种数据库的数百个源代码（如 ICD-9 和 ICD-10 代码）。

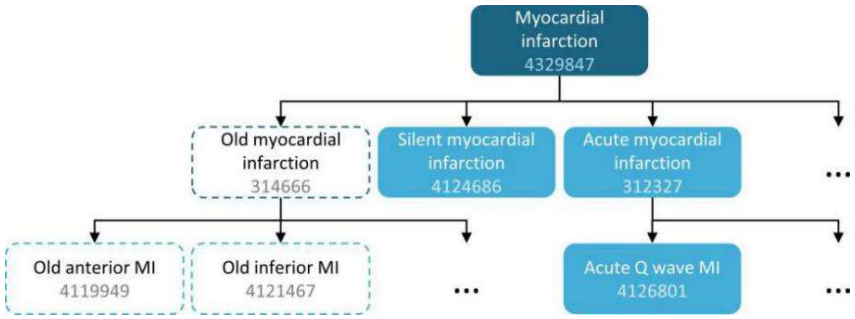


图 10.2 包括“心肌梗塞”（及子集）但不包括“旧心肌梗塞” （概念集及子集）的概念集

10.4 概率队列定义

基于规则的队列定义是队列定义中一种广泛使用的方法。然而，建立必要的专家共识来建立研究队列可能非常耗时。基于概率的队列设计是一种替代方法，可加快队列属性的选择过程。在这种方法中，有监督的机器学习让表型算法从一组有标记的示例（案例）中学习队列成员的相关属性。然后，可以使用该算法来更好的确定表型的特征定义，以及当选择修改表型标准时总体研究中准确定性之间的权衡。

将这种方法应用于 CDM 数据的一个示例是 R 包 APHRODITE（用于观察型定义、识别、训练和评估的自动表型程序）。该包提供了一个队列构建的框架，结合了从不完全标记的数据中机器学习的能力（Banda et al. 2017）。

10.5 队列建立的有效性

建立队列时，你应该考虑一下哪个对你更重要：找到所有合格的患者？还是只找确定患病的患者？

构建队列的策略取决于专家对疾病定义的临床严格程度的共识。也就是说，队列设计取决于你要回答的研究问题。你可以选择建立一个队列定义，该定义使用你所能获得的一切信息，使用最低标准，以便您可以在 OHDSI 网站之间共享它，或者结合以上两者。最终，研究人员可以自行决定严格程度的阈值来研究感兴趣的队列。

如本章开头所提到的，队列定义是尝试从记录的数据中推断出我们想要观察的内容。这就引出了一个问题，我们这次尝试中取得了多少成就。通常，与金标准（如人工抽样审查）相比，基于规则的队列定义或概率算法是对队列建立的验证。在第 16 章（“临床有效性”）中对此进行了详细讨论。

10.5.1 OHDSI 金标准表型库

为了帮助OHDSI社区对现有队列定义和算法进行总体评估，成立了OHDSI金标准表型库(GSPL)

工作组。GSPL 工作组目的是通过基于规则和概率的方法来开发由 OHDSI 社区支持的表型库。GSPL 工作组使得 OHDSI 社区的成员可以找到、评估并利用经过社区验证的队列定义，来开展研究和其他活动。这些金标准定义将存在于库中，其条目将保留在特定的设计和评估标准中。有关 GSPL 的其他信息，请查阅 OHDSI 工作组页面。该工作组中的研究包括上一节中讨论的 APHRODITE （Banda et al.2017）和 PheValuator 工具（Swerdel, Hripcsak, and Ryan 2019），以及 OHDSI 网络研究中的共享电子病历和基因组学 eMERGE 表型库工作 （Hripcsak et al.2019）。如果你对表型审编感兴趣，请考虑为此工作组出一份力。

10.6 定义一个高血压队列

通过使用基于规则的方法组合队列的定义，我们开始实践定义队列的技能。在这个范例中，我们想要找到使用 ACEI 单药作为高血压一线治疗的患者。

基于这样的想法，我们现在来建立队列。在进行这项练习时，我们建立队列的逻辑类似于构建标准消耗表。图 10.3 展示了建立这个队列的逻辑框架。

您可以在 ATLAS 用户界面中构建一个队列，也可以直接针对 CDM 编写查询代码。我们将在本章中简要讨论这两个方法。

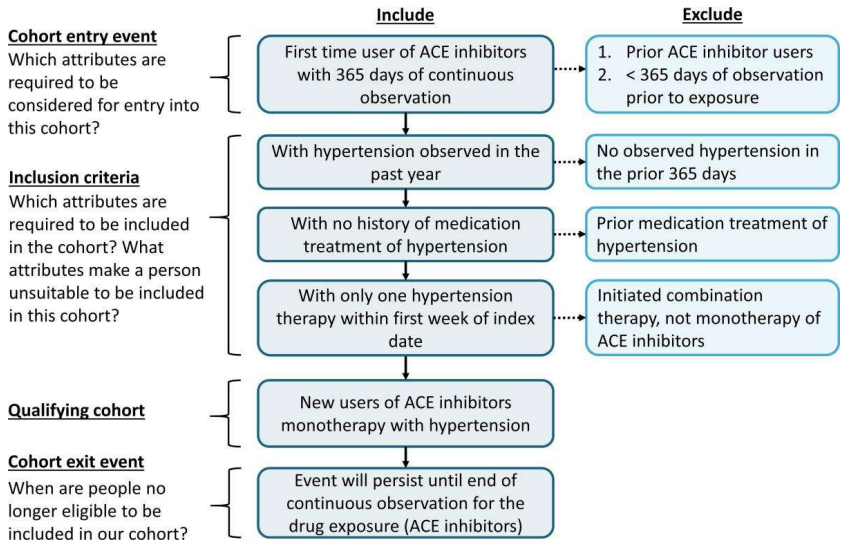


图 10.3：拟建队列的逻辑示意图

10.7 使用 ATLAS 建立队列

在 ATLAS 中首先单击 Cohort Definitions 模块。加载后，点击“新建队列”。下一个页面是空白的队列定义。图 10.4 显示了页面上的内容。

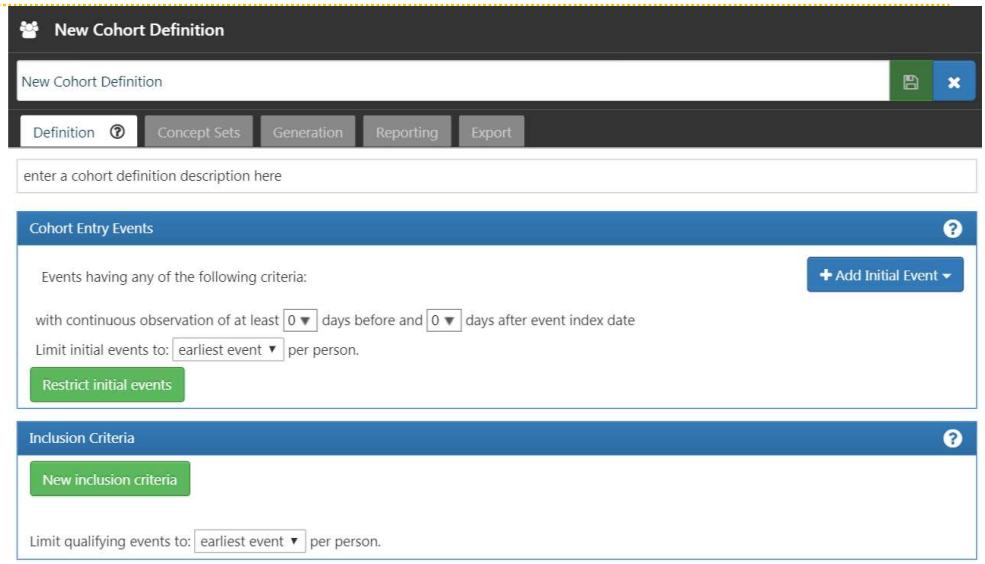


图 10.4: 新队列定义

在开始定义之前，建议将队列的名称从“新队列定义”更改为你对此队列的唯一名称。选择一个名字，如“ACEI 单药一线治疗高血压的新患者”。选好名字后，点击保存队列。

ATLAS 不允许两个队列有相同名字。如果您选择的名称已被另一个 ATLAS 队列使用，ATLAS 将弹出的错误消息。

10.7.1 初始事件的标准

继续定义初始队列事件。点击“添加初始事件”。这时需要选择围绕哪个域构建标准。你可能会问，“我怎么知道哪个域是初始队列事件?”让我们先说明这一点。如图 10.5 所示，ATLAS 在每个标准下面提供了相应描述。

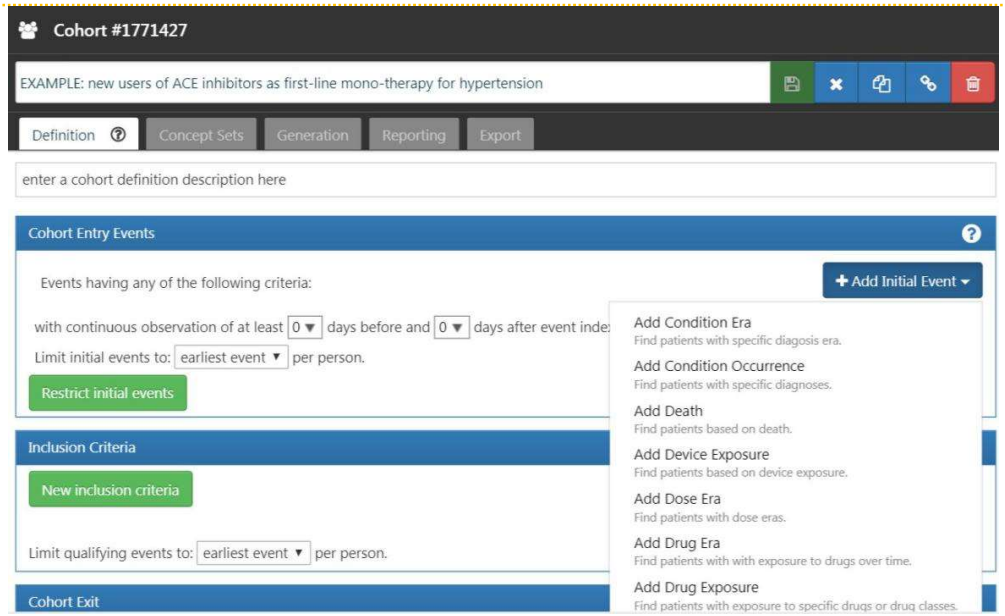


图 10.5：添加初始事件

如图 10.5 所示，ATLAS 在每个标准下面提供了相应描述。如果我们构建一个基于 CONDITION_OCCURRENCE（疾病状况）的标准，相应的问题会是查找有特定诊断的病人。如果我们建立一个基于 DRUG_EXPOSURE（药物暴露）的标准，相应的问题则是寻找使用特定药物或药物类别的患者。由于我们想查找开始将 ACEI 单药治疗作为高血压一线治疗的患者，这里需要选择一个 DRUG_EXPOSURE（药物暴露）的标准。你可能会说，“但我们也关注高血压诊断”。没错，高血压是我们要建立的另一个标准。然而，队列起始日期是由 ACEI 治疗的起始时间决定的，因此这是初始事件。高血压诊断是附加合格标准。我们之后将返回到这个问题。点击“添加 DRUG_EXPOSURE（药物暴露）”。

页面将更新所选标准，但还没有完成。如图 10.6 所示，ATLAS 不知道我们在寻找什么药物。因此需要告诉 ATLAS 哪个概念集与 ACEI 相关。

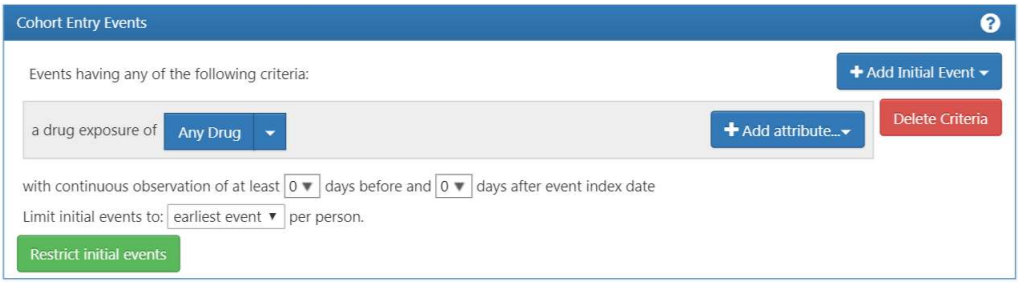


图 10.6：定义药物暴露

10.7.2 定义概念集

点击  打开对话框，从而检索一个概念集来定义 ACEI。

场景 1:您还没有构建概念集

针对所需标准，如果还没有概念集，那么首先要构建概念集。导航到“概念集”选项卡并单击“新概念集”，在队列定义中构建一个概念集。将概念集从“未命名的概念集”重命名为您选择的名称。在此基础上，您可以使用 Search 模块寻找 ACEI 相关的临床概念(图 10.7)。

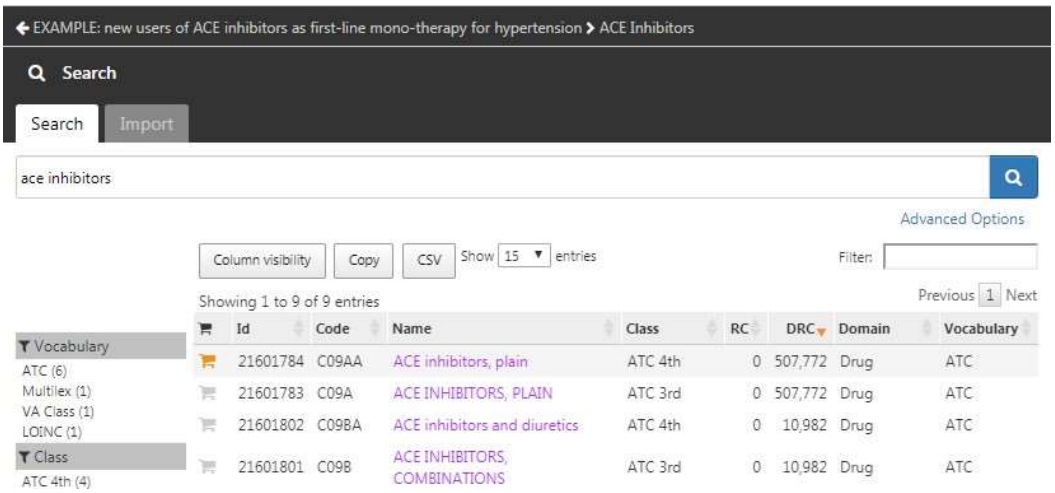


图 10.7：搜索词汇—— ACEI

找到需要的药物暴露概念后，你可以通过单击 选择概念。使用图 10.7 左上角的左箭头返回到队列定义。参考第 5 章(标准化词汇表)，了解如何导览这些词汇表，找到目标临床概念。

图 10.8 显示概念集表达式。我们选择了所有我们感兴趣的 ACEI 成分，并包括了它们的所有子概念，因此包括了所有含有这些成分的药物。点击“包含的概念”来查看这个表达式所隐含的所有 21,536 个概念，或点击“包含的源代码”来查看所隐含的各种编码系统中的所有源代码。

场景 2:您已经构建了一个概念集

如果你已经创建了概念集并保存于 ATLAS 中，点击“导入概念集”。对话框提示您在 ATLAS 概念集存储库中查找概念，如图 10.9 所示。示例图中，用户正在检索存储在 ATLAS 中的概念集。用户输入这个概念的名称，在右边的搜索中输入“ACEI”。此时概念集列表缩短为只有名称匹配的概念。接着，单击概念集的对应行来选择它。(注意:一旦选择了一个概念集，对话框就会消失。）当 Any Drug（任何药物） 的方框更新为你所选的概念集名称时，操作成功。

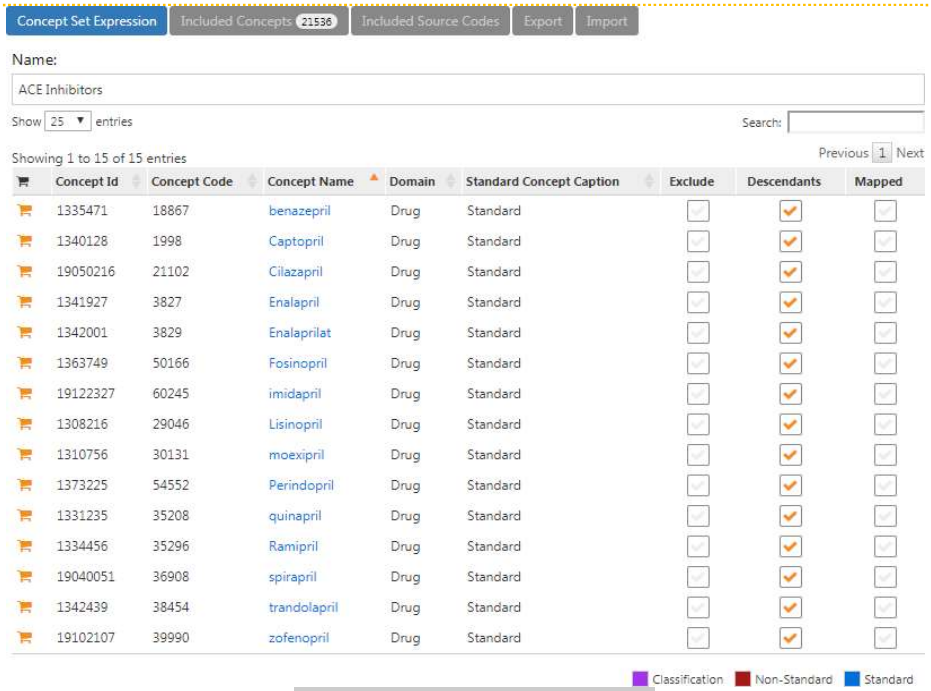


图 10.8：包含 ACEI 药物的概念集

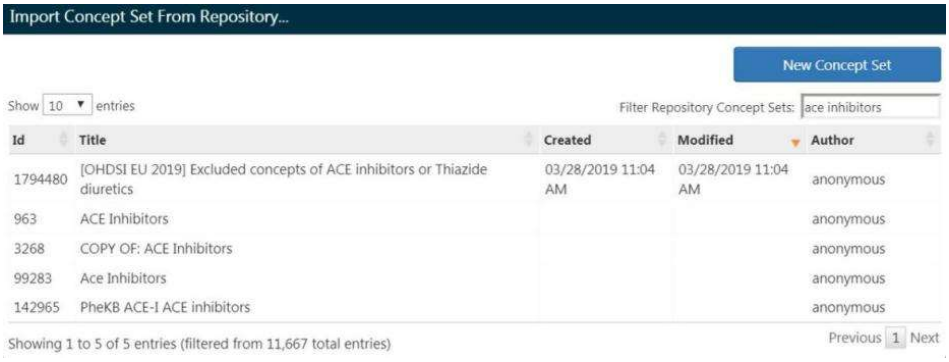


图 10.9：从 ATLAS 储存库导入一个概念集

10.7.3 附加的初始事件标准

目前已添加了一个概念集，但还没有完成。你的问题是寻找 ACEI 的新患者，或在病史中首次接触 ACEI 的人群。意即在患者记录中首次出现 ACEI。要指定此属性，单击“+Add attribute (添加属性)”。选择“添加第一个暴露条件”。注意，您可以指定构建标准的其他属性。可以指定事件发生时的年龄、发生日期、性别或与药物相关的其他属性。对于每个域，可选择标准不同。

接着窗口将自动关闭。一旦选中，这个附加属性将显示在与初始标准相同的框中(参见图 10.10)。

目前 ATLAS 的设计可能会让一些用户感到困惑。✖ 并不是意味着“否定或错误”。✖ 是一个可操作的特性，允许用户删除标准。如果单击，此条件将消失。因此，要保留 ✖ 符号，这个标准才是有效的。

现在您已经构建了一个符合标准的初始事件。为了确保抓取到第一个观察到的药物暴露，您将需要添加一个回望窗口，以确保查看足够多的患者病史，了解最先发生的情况。如果观察时间太短，病人可能在其他时间段也有同样的暴露事件。这一点无法控制，但我们可以规定一个在指标日期之前的最短时间段，在这段时间里患者数据依然符合条件。可以通过调整连续观察下拉菜单做到这点，也可单击该框并输入数值。比如我们需要在初始事件之前的 365 天连续观测。将观测周期更新为：连续观测 365 天，如图 10.10 所示。这个回望窗口由研究小组自行判断，在不同队列中可以有不同选择。这样就能尽可能地在最短的时间中确保我们抓取到第一个记录。这个标准是指标日期之前的历史时间，不涉及指标日期之后的时间。因此，我们设置指标事件后 0 天。我们的标准是首次使用 ACEI。因此，我们将初始事件限制为每个患者的 “最早事件”。

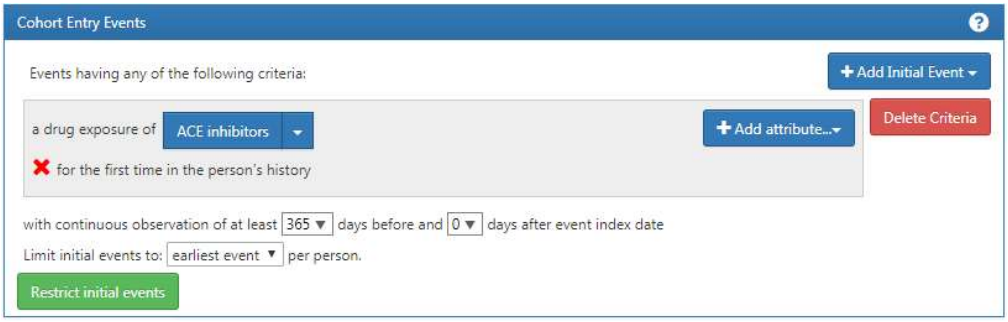


图 10.10：在索引日期之前设置所需的连续观察。

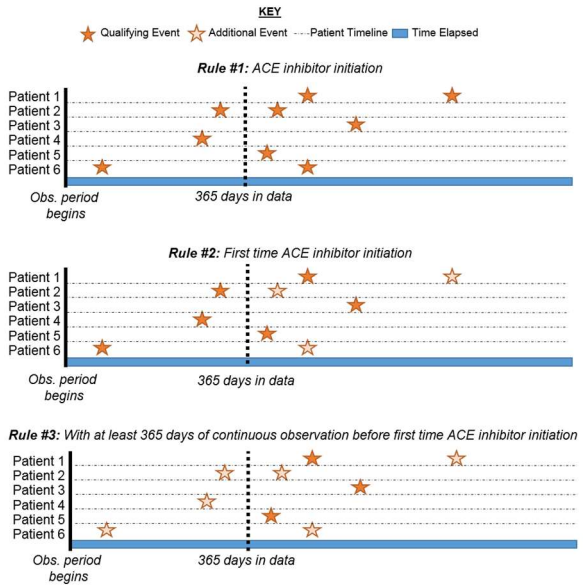


图 10.11:采纳不同标准下病人的入选资格

在图 10.11 中，每条线代表一个可能符合条件加入队列的患者。填充的星号表示患者满足指定标准的时间段。当加入额外标准时，一些星星的颜色变浅，表示这些病人的其它记录符合标准，但另有一些记录发生时间更早。设置完最后一个标准时，我们得到的是一个叠加筛选结果：首次使用 ACEI 的患者，且在首次使用前有 365 天的观察时间（即这 365 天中无 ACEI 使用事件译者注）。逻辑上，限制初始事件是多余的，但这有助于在每个选择中保持明确的逻辑。当你建立自己的队列时，可以参加 OHDSI 论坛的 Researchers 部分，获得关于如何构建队列逻辑的意见和想法。

10.7.4 入选标准

指定了一个队列进入事件之后，可以从这两处选项继续添加额外的限制条件：“Restrict initial events（限制初始事件）”和“New inclusion criteria（新纳入标准）”。这两个选项之间的根本区别在于 ATLAS 返回的临时信息。如果选择“Restrict initial events（限制初始事件）”，将新增限制标准添加到队列进入事件框中，选择在 ATLAS 中生成计数时，ATLAS 只返回满足所有这些标准的人数。如果选择将标准添加到“New inclusion criteria（新纳入标准）”中，您将得到一个损耗表，显示应用新增的纳入标准排除了多少患者。强烈建议使用纳入标准的方法，这样您就可以了解每个规则对队列定义总体完成的影响。你可能会发现某些纳入标准严重限制了最终进入队列的人数。你可以选择放宽这一标准，以获得更大的队列。这最终由定义该队列的专家们共同决定。

点击“New inclusion criteria（新纳入标准）”，添加关于该入组标准的后续逻辑。这个模块的工作方式与上文讨论的构建队列标准方法相同。你可以指定标准，添加特定的属性。我们的第一个附加标准纳入患者：*在指标日期后 365 天到 0 天之间至少发生 1 例高血压(首次使用 ACEI)*。点击“New inclusion criteria（新纳入标准）”添加。给所设标准命名，可以描述一下你的查找目标。目的是帮助你回忆所建立的内容，这不会影响你所定义的队列的完整性。

注释了这个新标准后，点击“+Add criteria to group（添加标准到组）”按钮来构建实际标准。该按钮功能类似于“添加初始事件”，只是不再指定初始事件。可以向其中添加多个条件——这也是它指定“添加标准到组”的原因。例如，查找疾病的方法有很多种(如，查找 CONDITION_OCCURRENCE 的逻辑、查找 DRUG_EXPOSURE 作为代理逻辑、查找 MEASUREMENT 作为代理逻辑)，这些是独立的域，需要不同的标准，但可以组合成一个标准来查找。按照这种方法，查找高血压的诊断，我们“Add condition occurrence（添加疾病状况）”；与建立初始事件类似的步骤一样，我们附加概念集；我们还要指定事件开始于指标日期(第一次使用 ACEI 的时间)之前的 365 天和之后的 0 天。参考图 10.12 检查逻辑。

接着添加查找患者的另一个标准：*在指定开始日期之前的所有时间到前 1 天，高血压药物的出现次数正好为 0 (在 ACEI 之前没有抗高血压药物暴露)*。这个过程和之前一样，首先点击“New inclusion criteria（新纳入标准）”，添加该标准的注释，然后点击“+Add criteria to group（添加标准到组）”。这是一个药物暴露，因此单击“Add Drug Exposure（添加药物暴露）”，附加一个高血压药物的概念集，并确定指标日期及范围(图中所示)。一定要确认您选择了正好 0 个事件。参考图 10.13 检查逻辑。

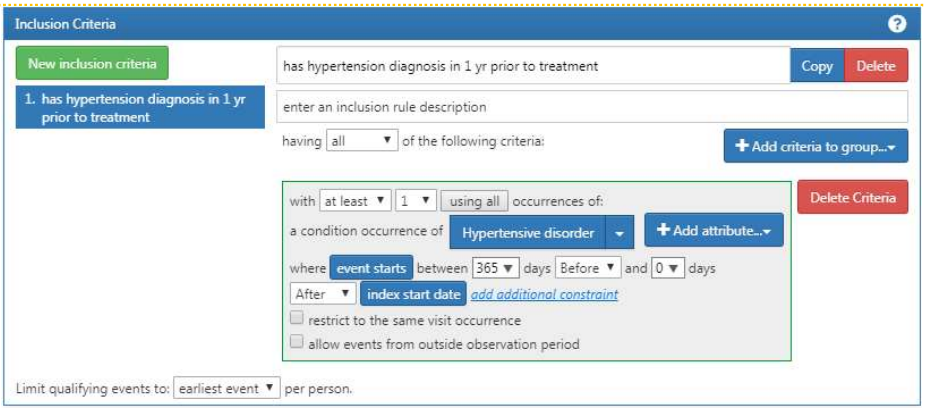


图 10.12：附加的纳入标准 1

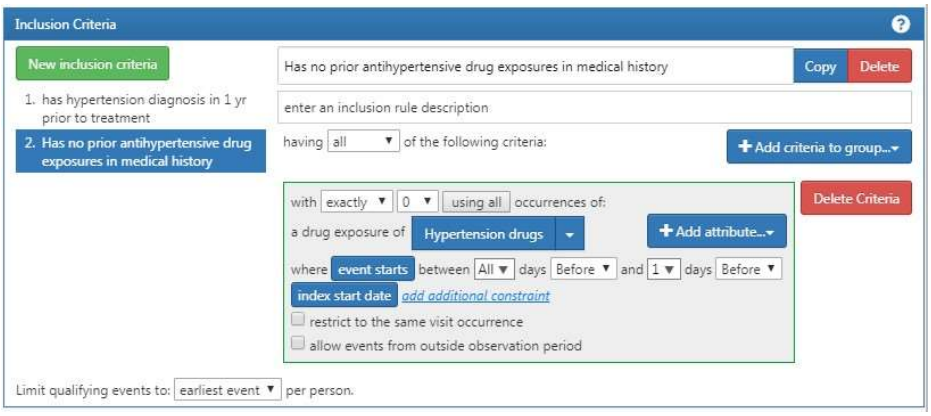


图 10.13：附加的纳入标准 2

您可能会问为什么“没有出现”编码为“恰好出现 0”。这是 ATLAS 处理信息的微妙之处。ATLAS 只使用包含标准。必须使用逻辑运算符来表达缺少特定的属性，比如：“Exactly 0（为 0)”。慢慢地，您将对 ATLAS 标准中可用的逻辑操作符更加熟悉。

最后，需添加另一个标准来查找患者：在指标事件开始日期前 0 天到后 7 天之间正好出现 1 次高血压药物处方事件，且只有一种抗高血压药物(为 ACEI)。和之前一样，首先点击“New inclusion criteria”按钮，添加注释，点击“+Add criteria to group”。这是一个 Drug_Era(药物时间)，点击“Add Drug Era（添加药物时间)”，附上一个高血压药物的概念集，设定指标日期前 0 天和后 7 天。参考图 10.14 检查逻辑。

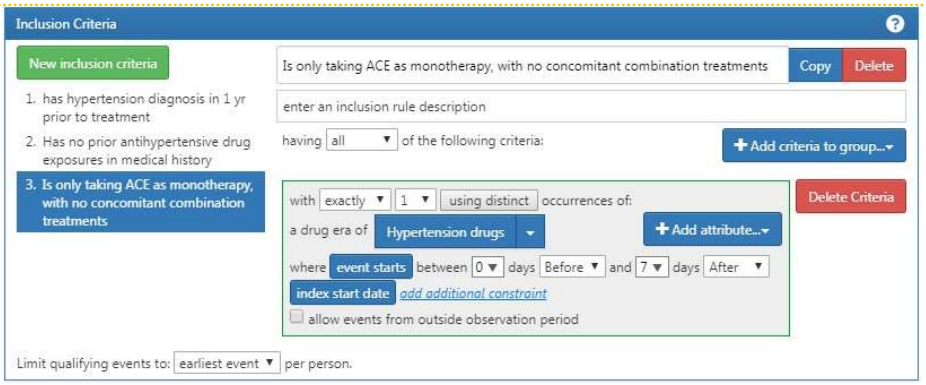


图 10.14：附加入选标准 3

10.7.5 队列排除标准

我们已经添加了所有的纳入标准。现在确定队列排除标准。你会问自己“什么时候患者不再有资格被纳入队列?”在这个队列中，我们跟踪新的药物暴露患者，所以需要考察药物暴露的连续观察时间。因此，排除标准可规定药物连续暴露的完整性：如果随后出现药物暴露中断，患者将在此时退出队列。因为我们无法确定在药物暴露中断期间患者发生了什么。我们还可以在持续性窗口上设置标准，以指定药物暴露之间允许的时间差。比如，领导这项研究的专家们决定，在认定持续暴露的时间段时，允许最多 30 天的暴露记录间隔。

为什么允许间隔? 在一些数据集中，我们只看到部分临床交互。尤其是药物暴露，一次处方配发可以覆盖一定的时间，因此允许药物暴露中一定的时间间隔。逻辑上病人依然在服用初始药物，因为单次药物配发覆盖超过一天。

通过选择 Event（事件）将持续至“end of a continuous drug exposure（持续药物暴露终止)”来配置它。然后添加我们的持续性窗口“允许最多 30 天”，并附加“ACEI”的概念集。根据图 10.15 检查逻辑。

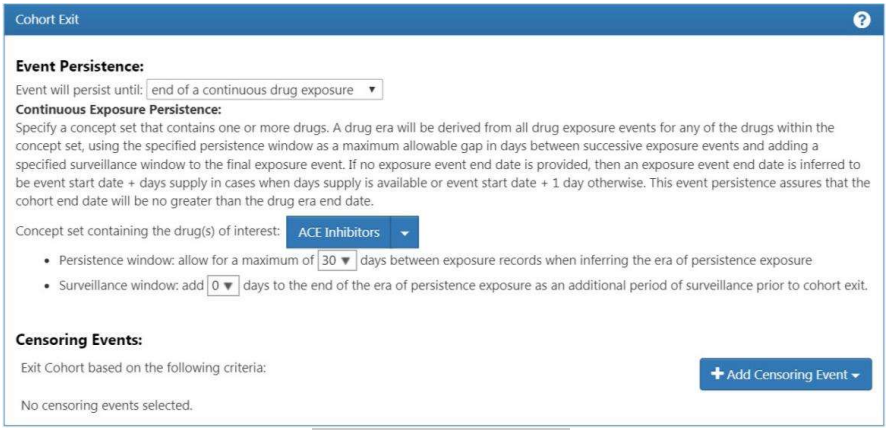


图 10.15：队列排除标准

在这个队列中，没有其他的审查事件（censoring events 选项)。不过你可能在构建其他队列需

要指定此条件，届时按照类似方法将其他属性添加到队列定义中即可。现已成功地创建了您的队列。务必按下  按钮保存。恭喜你！在 OHDSI 工具中，建立队列是回答问题的最重要的组成部分。现在使用“Export”选项卡将队列定义以 SQL 代码或 JSON 文件的形式共享给其他协作者，以加载到 ATLAS 中。

10.8 用 SQL 构建队列

这里我们将描述如何使用 SQL 和 R 创建相同的群组。在第九章提到，OHDSI 提供了两个 R 包，分别为 SqlRender 和 DatabaseConnector；两个 R 包一起加载，允许编写 SQL 代码，且自动翻译，并对各种数据库平台运行。为了表达清楚，我们将把 SQL 分成几个块，每个块生成一个临时表，可在下一个块中使用。这可能不是计算效率最高的方法，但比一个很长的语句更容易阅读。

10.8.1 连接到数据库

需要告诉 R 如何连接到服务器。我们使用 DatabaseConnector 包，它提供了一个名为 createConnectionDetails 的函数。键入?createConnectionDetails 用于不同数据库所需的特定数据库管理系统(DBMS)。例如，可以使用以下代码连接到 PostgreSQL 数据库：

```
library(CohortMethod)
connDetails <- createConnectionDetails(dbms = "postgresql",
                                       server = "localhost/ohdsi",
                                       user = "joe",
                                       password = "supersecret")

cdmDbSchema <- "my_cdm_data"
cohortDbSchema <- "scratch"
cohortTable <- "my_cohorts"
```

最后三行定义 cdmDbSchema, cohortDbSchema, 和 cohortTable 变量。我们稍后将使用这些来告诉 R CDM 格式的数据位于何处，以及需要在何处创建感兴趣的队列。注意对于 Microsoft SQL Server，数据库模式需同时指定数据库和模式，例如 cdmDbSchema< - “my_cdm_data.dbo”。

10.8.2 指定概念

为了使其可读，我们将定义 R 中需要的概念 id，并传递给 SQL：

```
aceI <- c(1308216, 1310756, 1331235, 1334456, 1335471, 1340128, 1341927,
          1342439, 1363749, 1373225)
hypertension <- 316866
allHtDrugs <- c(904542, 907013, 932745, 942350, 956874, 970250, 974166,
                978555, 991382, 1305447, 1307046, 1307863, 1308216,
                1308842, 1309068, 1309799, 1310756, 1313200, 1314002,
                1314577, 1317640, 1317967, 1318137, 1318853, 1319880,
                1319998, 1322081, 1326012, 1327978, 1328165, 1331235,
                1332418, 1334456, 1335471, 1338005, 1340128, 1341238,
                1341927, 1342439, 1344965, 1345858, 1346686, 1346823,
                1347384, 1350489, 1351557, 1353766, 1353776, 1363053,
                1363749, 1367500, 1373225, 1373928, 1386957, 1395058,
                1398937, 40226742, 40235485)
```

10.8.3 查找首次使用

首先查找每个患者 ACEI 的首次使用：

```
conn <- connect(connectionDetails)

sql <- "SELECT person_id AS subject_id,
  MIN(drug_exposure_start_date) AS cohort_start_date
INTO #first_use
FROM @cdm_db_schema.drug_exposure
INNER JOIN @cdm_db_schema.concept_ancestor
  ON descendant_concept_id = drug_concept_id WHERE
ancestor_concept_id IN (@ace_i)
GROUP BY person_id;"

renderTranslateExecuteSql(conn,
                          sql,
                          cdm_db_schema =cdmDbSchema,
                          ace_i = aceI)
```

注意，我们将 DRUG_EXPOSURE 表连接到 CONCEPT_ANCESTOR 表，以查找包含 ACEI 的所有药物。

10.8.4 要求 365 天的预先观察

接下来，我们需要连接 OBSERVATION_PERIOD，定义预先连续观测 365 天：

```
sql <- "SELECT
  subject_id,
  cohort_start_date
INTO
#has_prior_obs
FROM #first_use
INNER JOIN @cdm_db_schema.observation_period
  ON subject_id = person_id
    AND observation_period_start_date <= cohort_start_date
    AND observation_period_end_date >= cohort_start_date
WHERE DATEADD(DAY, 365, observation_period_start_date) < cohort_start_date;"

renderTranslateExecuteSql(conn, sql, cdm_db_schema = cdmDbSchema)
```

10.8.5 要求先前有高血压诊断

```
sql <- "SELECT DISTINCT
  subject_id,
  cohort_start_date
INTO #has_ht
FROM #has_prior_obs
INNER JOIN
  @cdm_db_schema.condition_occurrence ON
  subject_id = person_id
    AND condition_start_date <= cohort_start_date
    AND condition_start_date >= DATEADD(DAY, -365, cohort_start_date)
INNER JOIN @cdm_db_schema.concept_ancestor
  ON descendant_concept_id =
condition_concept_id WHERE
ancestor_concept_id = @hypertension;"
renderTranslateExecuteSql(conn,sql,
                          cdm_db_schema =cdmDbSchema,
                          hypertension =hypertension)
 renderTranslateExecuteSql(conn,sql,
                          cdm_db_schema =cdmDbSchema,
                          hypertension =hypertension)
```

我们要求在 365 天内先前有高血压诊断：

注意，我们 SELECT DISTINCT（选择不重复的），因为如果一个人在过去有多个高血压诊断，我们可能创建重复的队列条目。

10.8.6 先前无其他治疗

我们要求先前无其他高血压治疗：

```
sql <- "SELECT subject_id,
  cohort_start_date
INTO #no_prior_ht_drugs
FROM #has_htEFT JOIN

 ( SELECT *
   FROM @cdm_db_schema.drug_exposure
   INNER JOIN @cdm_db_schema.concept_ancestor
     ON descendant_concept_id = drug_concept_id
   WHERE ancestor_concept_id IN (@all_ht_drugs)
 ) ht_drugs
   ON subject_id = person_id
     AND drug_exposure_start_date < cohort_start_date
 WHERE person_id IS NULL;"

 renderTranslateExecuteSql(conn,
                           sql,
                           cdm_db_schema =cdmDbSchema,
                           all_ht_drugs = allHtDrugs)
```

注意，我们使用了 left join，并且只允许 person_id(来自 DRUG_EXPOSURE 表)为空的行，即没有找到匹配的记录。

10.8.7 单药治疗

我们要求在队列入组的前 7 天，高血压治疗用药只有一种：

```
sql <- "SELECT subject_id,
  cohort_start_date
INTO #monotherapy
FROM #no_prior_ht_drugs

INNER JOIN @cdm_db_schema.drug_exposure
  ON subject_id = person_id
    AND drug_exposure_start_date >= cohort_start_date
    AND drug_exposure_start_date <= DATEADD(DAY, 7, cohort_start_date)
INNER JOIN @cdm_db_schema.concept_ancestor
  ON descendant_concept_id = drug_concept_id
WHERE ancestor_concept_id IN (@all_ht_drugs)
GROUP BY subject_id,
  cohort_start_date
HAVING COUNT(*) = 1;"

renderTranslateExecuteSql(conn,sql,
                          cdm_db_schema = cdmDbSchema,
                          all_ht_drugs = allHtDrugs)
```

10.8.8 退出队列

除了队列结束日期外，队列的其他定义已全部完成。根据定义，药物暴露结束时患者退出队列，允许间隔最多 30 天。因此不仅需要考虑第一次药物暴露，还要考虑随后的 ACEI 药物暴露。将随后的暴露组合成区间的 SQL 可能非常复杂。幸运的是，标准代码已经被定义，可以有效地用于创建区间。(这段代码由 Chris Knoll 编写，在 OHDSI 中经常被称为“魔法”)。我们首先创建一个临时表包含所有我们希望合并的药物暴露：

```
sql <- "
  SELECT person_id,
    CAST(1 AS INT) AS concept_id,
    drug_exposure_start_date AS exposure_start_date,
    drug_exposure_end_date AS exposure_end_date
  INTO #exposure
  FROM @cdm_db_schema.drug_exposure
  INNER JOIN @cdm_db_schema.concept_ancestor
    ON descendant_concept_id = drug_concept_id
  WHERE ancestor_concept_id IN (@ace_i);"
renderTranslateExecuteSql(conn,sql,
                          cdm_db_schema = cdmDbSchema,
                          ace_i = aceI)
```

然后我们运行按顺序合并暴露的标准代码：

```
sql <- "
SELECT ends.person_id AS subject_id,
       ends.concept_id AS cohort_definition_id,
   MIN(exposure_start_date) AS cohort_start_date,
   ends.era_end_date AS cohort_end_date
 INTO #exposure_era
FROM (
   SELECT exposure.person_id,
       exposure.concept_id,
       exposure.exposure_start_date,
       MIN(events.end_date) AS era_end_date
   FROM #exposure exposure
   JOIN (
--cteEndDates
       SELECT person_id,
           concept_id,
           DATEADD(DAY, -1 * @max_gap, event_date) AS end_date
       FROM (
           SELECT person_id,
           concept_id,
           event_date,
           event_type,
           MAX(start_ordinal) OVER (
               PARTITION BY person_id ,concept_id ORDER BY event_date,
               event_type ROWS UNBOUNDED PRECEDING
           ) AS start_ordinal,
           ROW_NUMBER() OVER (
               PARTITION BY person_id, concept_id ORDER BY event_date,
           event_type
           ) AS overall_ord
   -- select the start dates, assigning a row number to each
           SELECT person_id,
                       concept_id,


                 exposure_start_date AS event_date,
                 0 AS event_type,
                 ROW_NUMBER() OVER (
                   PARTITION BY person_id, concept_id ORDER BY
               exposure_start_date
                        ) AS start_ordinal
               FROM #exposureexposure

               UNION ALL
   -- add the end dates with NULL as the row number, padding the end dates by
   -- @max_gap to allow a grace period for overlapping ranges.

               SELECT person_id, concept_id,
                   DATEADD(day, @max_gap, exposure_end_date),
                   1 AS event_type,
                   NULL
               FROM #exposure exposure

       ) rawdata
   ) events
  WHERE 2 * events.start_ordinal - events.overall_ord = 0
```

```
    ) events
    ON exposure.person_id =events.person_id

          ) rawdata
      ) events
    WHERE 2 * events.start_ordinal - events.overall_ord = 0
    ) events
    ON exposure.person_id =events.person_id
        AND exposure.concept_id =events.concept_id
        AND events.end_date >= exposure.exposure_end_date
    GROUP BY exposure.person_id,
        exposure.concept_id,
        exposure.exposure_start_date
    ) ends
  GROUP BY ends.person_id,
    concept_id,
    ends.era_end_date;"

  renderTranslateExecuteSql(conn,sql,
                            cdm_db_schema =cdmDbSchema,
                            max_gap = 30)
```

这段代码合并了所有的后续药物暴露，用 max_gap 参数定义暴露之间的允许间隔。产生的药物暴露时段记录在#exposure_era 临时表格中。接着，将 ACEI 暴露时段加入到原始队列中，使用区间结束日期作为队列结束日期：

```
sql <- "SELECT ee.subject_id,
  CAST(1 AS INT) AS cohort_definition_id,
  ee.cohort_start_date,
  ee.cohort_end_date
INTO @cohort_db_schema.@cohort_table
FROM #monotherapy mt
INNER JOIN #exposure_era ee
  ON mt.subject_id = ee.subject_id
    AND mt.cohort_start_date =ee.cohort_start_date;"

renderTranslateExecuteSql(conn,sql,
                          cohort_db_schema = cohortDbSchema,
                          cohort_table =cohortTable)
```

这里，我们把最后形成的队列存储在之前定义的模式（schema）和表（table）中。分配队列定义 ID 为 1，以区别于将来存储在同一表中的其他队列。

10.8.9 清理

最后，建议清理创建的临时表，并断开与数据库服务器的连接：

```
sql <- "TRUNCATE TABLE #first_use;
DROP TABLE #first_use;

TRUNCATE TABLE #has_prior_obs;
DROP TABLE #has_prior_obs;

TRUNCATE TABLE #has_ht;
DROP TABLE #has_ht;

TRUNCATE TABLE #no_prior_ht_drugs;
DROP TABLE #no_prior_ht_drugs;

TRUNCATE TABLE #monotherapy;
DROP TABLE #monotherapy;

TRUNCATE TABLE #exposure;
DROP TABLE #exposure;

TRUNCATE TABLE #exposure_era;
DROP TABLE #exposure_era;"
renderTranslateExecuteSql(conn, sql)
disconnect(conn)
```

10.9 总结

-

- SQL（结构化查询语言）是用于查询符合通用数据模型（CDM）的数据库的
- 队列是在一段时间内满足一个或者多个入选标准的一组人。
- 队列定义是用来识别某一个特定队列的逻辑性的描述。
- 在 OHDSI 分析工具中，队列被用来（并反复使用）定义感兴趣的风险和结果。
- 建立队列的主要方法有两种：基于规则的方法和基于概率的方法。
- 可以在 ATLAS 或者使用 SQL 创建基于规则的队列定义。

10.10 练习

先决条件

对于第一个练习，你需要访问 ATLAS。

你可以访问网页 http://atlas-demo.ohdsi.org ，或者访问其他你有权限的网页。

练习 10.1 使用 ATLAS 来建立符合以下条件的队列

- 双氯芬酸的新使用者
- 年龄为 16 岁或者以上
- 在药物暴露之前有至少 365 天的联系观测记录
- 先前没有用过任何的非甾体抗炎药

- 先前没有癌症的诊断
- 队列推出条件为连续用药结束（允许 30 天的间隔期）

先决条件

对于第二个练习，我们需要按照第 8.4.5 章节中的说明安装 R，R-studio，和 Java。同时还需要安装 SqlRender，DatabaseConnector，和 Eunomia 包，可以使用以下方法进行安装：

```
install.packages(c("SqlRender", "DatabaseConnector", "devtools"))
devtools::install_github("ohdsi/Eunomia", ref = "v1.0.0")
```

Eunomia 包提供了一个模拟的 CDM 数据集，该数据集将在你的本地 R 中运行。可以使用以下方法连接数据集：

```
connectionDetails <- Eunomia::getEunomiaConnectionDetails()
```

使用的 CDM 数据库是“main”。

练习 10.2 使用 R 和 SQL 在现有的 COHORT 表中建立满足以下条件的急性心肌梗塞队列(AMI)：

- 诊断为心肌梗塞 (概念概念 4329847 “心肌梗塞” 及其所有子集，概念但不包括 314666 “旧心肌梗塞” 及其所有子集)
- 在住院或者急诊就诊期间（概念概念 9201，9203，262 分别代表 “住院就诊”，“急诊室就诊”，“急诊室和住院就诊”）

答案可以在附录 E.6 找到。

参考文献

1. Banda, J. M., Y. Halpern, D. Sontag, and N. H. Shah. 2017. “Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network.” *AMIA Jt Summits Transl Sci Proc* 2017: 48–57.
2. Hripcsak, G., N. Shang, P. L. Peissig, L. V. Rasmussen, C. Liu, B. Benoit, R. J. Carroll, et al. 2019. “Facilitating phenotype transfer using a common data model.” *J Biomed Inform*, July, 103253.
3. Swerdel, J. N., G. Hripcsak, and P. B. Ryan. 2019. “PheValuator: Development and Evaluation of a Phenotype Algorithm Evaluator.” J Biomed Inform, July, 103258.
4. Swerdel, J. N., G. Hripcsak, and P. B. Ryan. 2019. “PheValuator: Development and Evaluation of a Phenotype Algorithm Evaluator.” *J Biomed Inform*, July, 103258.
5. Kahn, Michael G., Jeffrey S. Brown, Alein T. Chun, Bruce N. Davidson, Daniella Meeker, P. B. Ryan, Lisa M. Schilling, Nicole G. Weiskopf, Andrew E. Williams, and Meredith Nahm Zozus. 2015. “Transparent Reporting of Data Quality in Distributed Data Networks.” *EGEMS (Washington, DC)* 3 (1): 1052. https://doi.org/10.13063/2327-9214.1052.
6. Kahn, Michael G, Marsha A Raebel, Jason M Glanz, Karen Riedlinger, and John F Steiner. 2012. “A Pragmatic Framework for Single-Site and Multisite Data Quality Assessment in Electronic Health Record-Based Clinical Research.” *Medical Care* 50.
7. Liaw, Siaw-Teng, Alireza Rahimi, Pradeep Ray, Jane Taggart, Sarah Dennis, Simon de Lusignan, B Jalaludin, AET Yeo, and Amir Talaei-Khoei. 2013. “Towards an Ontology for Data Quality in Integrated Chronic Disease Management: A Realist Review of the Literature.” *International Journal of Medical Informatics* 82 (1): 10–24.
8. Madigan, D., P. B. Ryan, and M. Schuemie. 2013. “Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies.” *Ther Adv Drug Saf* 4 (2): 53–62.

9. Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. 2013. “Evaluating the impact of database heterogeneity on observational study results.” *Am. J. Epidemiol.* 178 (4): 645–51.
10. Schuemie, M. J., G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. 2016. “Robust empirical calibration of p-values using observational data.” *Stat Med* 35 (22): 3883–8.
11. Schuemie, M. 2018. “Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data.” *Proc. Natl. Acad. Sci. U.S.A.* 115 (11): 2571–7.
12. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. “Improving reproducibility by using high-throughput observational studies with empirical calibration.” *Philos Trans A Math Phys Eng Sci* 376 (2128).
13. Sherman, Rachel E, Steven A Anderson, Gerald J Dal Pan, Gerry W Gray, Thomas Gross, Nina L Hunter, Lisa LaVange, et al. 2016. “Real-World Evidence—What Is It and What Can It Tell Us.” *N Engl J Med* 375 (23): 2293–7.
14. Weiskopf, Nicole Gray, and Chunhua Weng. 2013. “Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research.” *Journal of the American Medical Informatics Association: JAMIA* 20 (1): 144–51. https://doi.org/10.1136/amiajnl-2011-000681.
15. Yoon, D., E. K. Ahn, M. Y. Park, S. Y. Cho, P. Ryan, M. J. Schuemie, D. Shin, H. Park, and R. W. Park. 2016. “Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research.” *Healthc Inform Res* 22 (1): 54–58.

---

54. https://github.com/OHDSI/Aphrodite

55. https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:gold-library-wg

术语概览

特征提取	FeatureExtraction	
机器学习应用	machine learning applications	
数据库管理系统	database management systems	DBMS
特征描述	Characterization	
索引后	post-index	
发病率	incidence proportion	
发生率	incidence rate	

第十一章　特征描述

章节负责人: Anthony Sena & Daniel Prieto-Alhambra

观察性健康数据库提供了一种有价值的、可以基于一系列特征来了解人群变异的资源。 通过使用描述性统计学方法来对人群进行特征描述是产生各种健康和疾病决定因素假设重要的第一步。 在这一章中，我们将介绍特征描述的方法:

- **数据库层级数据特征描述**: 提供一组最高级别的汇总统计数据，以了解整个数据库的数据概况。
- **队列特征描述:** 根据总体病史来描述一个人群。
- **治疗路径:** 描述一个人在一段时间内接受干预的顺序。
- **发病率:** 衡量一段时间内在基本人群中结果事件的发生比率。

除了数据库层级的特征描述，这些方法的目的是描述与被特指为索引日期的事件相关的人群。这个特征人群即第十章介绍的“队列”。该队列定义了相关人群中每个人的索引日期。使用索引日期作为锚点，我们将索引日期之前的时间定义为基线时间，索引日期和之后的所有时间称为索引后时间。

特征描述的用例包括自然病史、治疗有效性和质量改进。在这一章中，我们将介绍特征描述的方法。我们将使用一个高血压患者人群来演示如何使用 ATLAS 和 R 来执行这些特征描述任务。

11.1 数据库层级特征描述

在我们回答任何关于目标人群的特征描述问题之前，必须先了解我们打算使用的数据库的特征。数据库层级的特征描述意图从时间趋势和分布的角度来描述一个数据库的整体。对数据库的这种定量评估通常包括以下问题:

- 这个数据库中的总人数是多少?
- 人口的年龄分布情况如何?
- 这个数据库中的人被观察了多长时间?
- 随着时间的推移，有(治疗、状况、治疗操作等)记录 / 处方的人的比例是多少?

这些数据库层级的描述统计方法还可以帮助研究人员了解数据库中可能缺少的数据。第 15 章会进一步详细讨论数据质量。

11.2 队列特征描述

队列特征描述包括队列中人群的基线特征和索引后特征。OHDSI 通过对患者病史中所有病情、药物和器械暴露、操作和其他临床观察的特征描述，通过描述统计方法进行评估。我们还汇总了索引日期的队列成员的人口统计学资料。这种方法提供了目标队列的完整摘要。重要的是，这使在关注数据变化的同时全面探索队列特征成为可能，同时也让我们可以识别潜在的缺失值。

队列特征描述方法可用于个人水平的药物利用研究(DUS) 以估计使用某一治疗的患者人群中适应症和禁忌症的发生率。正如《加强流行病学观察研究报告》(STROBE)指南中详细阐述的那样，传播队列的特征描述信息所是一项观察型研究中推荐的最佳实践。 (Elm et al. 2008)

11.3 治疗路径

描述人群特征的另一种方法是描述索引后时间窗内的治疗顺序。例如，Hripcsak 等人(2016 年)利用 OHDSI 通用数据标准创建了描述 2 型糖尿病、高血压和抑郁症治疗路径的描述统计方法。通过标准化这种分析方法，Hripcsak 和同事们可以在 OHDSI 网络上运行相同的分析来描述目标人群的特征。

分析治疗路径的目的是总结诊断为某一特定病情的患者从首次处方/配药开始的治疗(事件)。在此案例中，描述的是患者被诊断为 2 型糖尿病，高血压和抑郁症后的治疗。每个患者的所有事件被汇总成一套统计摘要并针对每个状况和每个数据库进行可视化。

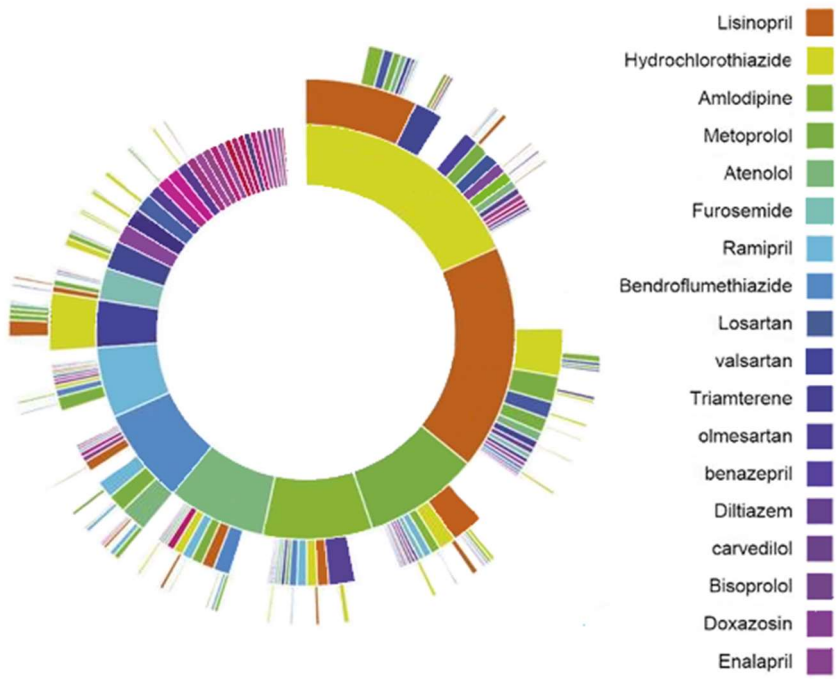


图 11.1: OHDSI 高血压治疗路径可视化旭日图

例如，图 11.1 表示正在开始接受高血压治疗的人群。中间的第一圈显示了基于一线治疗的人数比例。 在这个例子中，氢氯噻嗪是这个人群最常见的一线治疗。 从氢氯噻嗪部分延伸出来的箱子代表队列中记录的二线和三线疗法。

路径分析提供了人群中治疗利用度的重要证据。通过本分析，我们可以描述最通用的一线治疗利用情况，停止治疗的人群情况，换用其他治疗方案或增加原治疗强度的人群情况。 通过路径分析，Hripcsak 等人(2016 年)发现二甲双胍是最常用的糖尿病处方药，从而证实了美国临床内分泌专家协会关于糖尿病治疗路径的一线建议被普遍采用。此外，他们还注意到，10%的糖尿病患者、24%的高血压患者和 11%的抑郁症患者遵循的治疗路径在任何数据来源中都与其他人不同。

在经典的 DUS 术语中，治疗路径分析包括一些人群水平的 DUS 估计，如在特定人群中使用一种或多种药物的普遍程度，以及一些个人水平的 DUS，包括不同治疗方案的维持、变更情况。

11.4 发病率

发病率和比例是公共卫生中用于评估处于风险暴露期间(TAR)的人群新结局事件发生情况的统计指标。图 11.2 旨在显示单个人的发病率计算的组成部分:

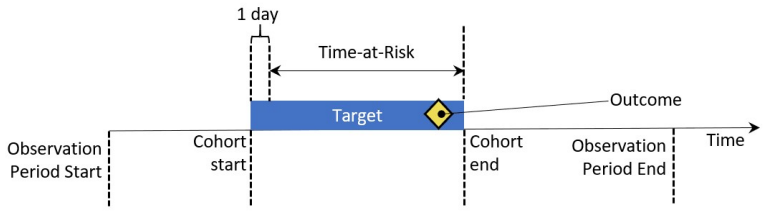


图 11.2: 个人水平的发病率计算要素视图。本例中，风险暴露期定义为队列开始后一天开始，队列结束时结束。

在图 11.2 中，一个人在数据中的观察时间由观察开始和结束时间表示。接下来，此人会因为满足一些标准在某一时间点进入和退出一个队列。风险暴露期窗口即我们想了解一个结果发生情况的时间段。如果结果发生于风险暴露期间，则将其计入结果的发生率。

计算发病率有两个指标:

$$Incidence\ Proportion = \frac{\#\ persons\ in\ cohort\ with\ new\ outcome\ during\ TAR}{\#\ persons\ in\ cohort\ with\ TAR}$$

发病比例提供了在风险暴露期间人群里，人均发生目标结果的衡量标准。 换句话说，这是在一个确定的时间范围内发生了目标结果的人在观察队列中的比例。

$$Incidence\ Rate = \frac{\#\ persons\ in\ cohort\ with\ new\ outcome\ during\ TAR}{person\ time\ at\ risk\ contributed\ by\ persons\ in\ cohort}$$

发病率是衡量人群在累计风险暴露期间内发生目标结果数量的指标。当一个人在风险暴露期间发生了目标结果，其对总人时的贡献在结果事件发生时即停止。累计风险暴露时间称为人时，以天、月或年表示。

发生比例和发病率在治疗上的应用就是典型的特定治疗在人群水平的 DUS。

11.5 高血压患者特征

根据世界卫生组织(WHO)关于高血压的全球简报(WHO，2013 年) ，及早发现、适当治疗和良好控制高血压在卫生和经济方面都有重大收益。世界卫生组织的简报概述了高血压并描述了不同国家的高血压疾病负担。世界卫生组织根据地理区域、社会经济阶层和性别对高血压进行统计描述。

观察型数据库提供了类似世界卫生组织描述高血压人群特征的方法。在本章的后续章节中，我们将探讨如何利用 ATLAS 和 R 探索数据库以了解其构成，并用于研究高血压人群。然后，我们将使用相同的工具来描述高血压人群的自然历史和治疗模式。

11.6 ATLAS 中的数据库 / 特征描述

这里我们介绍如何使用 ATLAS 中的数据源模块来了解 ACHILLES 创建的特征描述统计数据库，以发现与高血压患者相关的数据库级特征。首先点击 ATLAS 的左栏 Data Sources。在 ATLAS 中显示的第一个下拉列表中，选择要查看的数据库。接下来，使用数据库下方的下拉菜单开始查看报表。要做到这一点，需在 Report 的下拉菜单中选择“状况发生 Condition Occurrence”来显示数据库中所有状况的树形图：

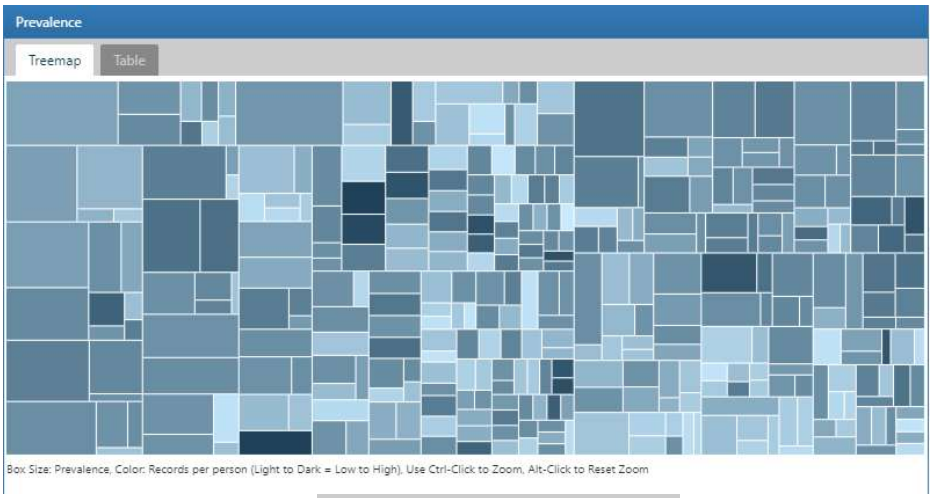
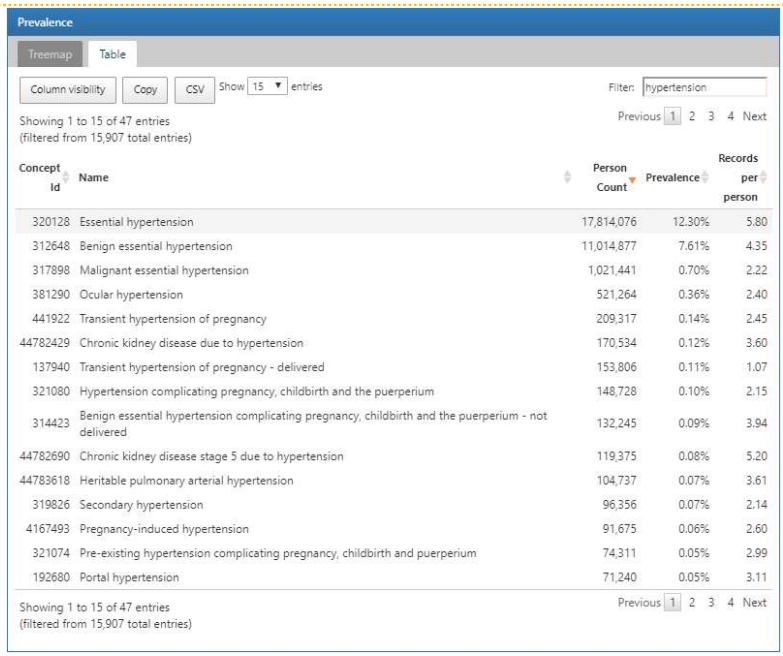


图 11.3: Atlas 数据来源: 状况发生树形图

要搜索特定的目标状况，单击 Table 选项卡可以显示数据库中状况的完整列表，包括人数、患病率和每人记录。使用顶部的过滤框，我们可以通过包含术语“高血压”的概念名称过滤表中的条目：



Prevalence

Treemap | Table

Column visibility | Copy | CSV | Show 15 entries | Filter: hypertension

Showing 1 to 15 of 47 entries
(filtered from 15,907 total entries)

Previous 1 2 3 4 Next

Concept Id	Name	Person Count	Prevalence	Records per person
320128	Essential hypertension	17,814,076	12.30%	5.80
312648	Benign essential hypertension	11,014,877	7.61%	4.35
317898	Malignant essential hypertension	1,021,441	0.70%	2.22
381290	Ocular hypertension	521,264	0.36%	2.40
441922	Transient hypertension of pregnancy	209,317	0.14%	2.45
44782429	Chronic kidney disease due to hypertension	170,534	0.12%	3.60
137940	Transient hypertension of pregnancy - delivered	153,806	0.11%	1.07
321080	Hypertension complicating pregnancy, childbirth and the puerperium	148,728	0.10%	2.15
314423	Benign essential hypertension complicating pregnancy, childbirth and the puerperium - not delivered	132,245	0.09%	3.94
44782690	Chronic kidney disease stage 5 due to hypertension	119,375	0.08%	5.20
44783618	Heritable pulmonary arterial hypertension	104,737	0.07%	3.61
319826	Secondary hypertension	96,356	0.07%	2.14
4167493	Pregnancy-induced hypertension	91,675	0.06%	2.60
321074	Pre-existing hypertension complicating pregnancy, childbirth and puerperium	74,311	0.05%	2.99
192680	Portal hypertension	71,240	0.05%	3.11

Showing 1 to 15 of 47 entries
(filtered from 15,907 total entries)

Previous 1 2 3 4 Next

图 11.4: Atlas 数据来源: 在概念名称检索到的包含“高血压”的状况

我们可以通过单击某一行来查看某一状况的详细报告。本例中我们会选择“原发性高血压”，以获取所选状况随时间的推移和按性别分布的患病率、按月份状况分布的患病率、状况记录的类型和首次诊断时的年龄:

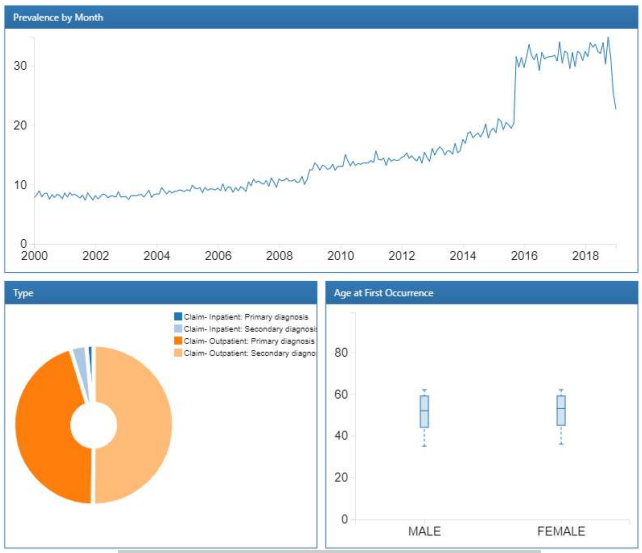


图 11.5: Atlas 数据来源: 原发性高血压详细报告

我们已经回顾了数据库中高血压概念的特征和随时间变化的趋势，我们也可以探索用于治疗高血压患者的药物。这个过程遵循相同的步骤(除了我们用 RxNorm 成分归纳的“药物疗程报告”来分析药

物特征)。一旦通过数据库级特征描述了解了目标事件，我们可以进一步建立队列来识别高血压患者的特征。

11.7 ATLAS 中的特征描述

在这里，我们演示如何使用 ATLAS 来执行大规模的队列特征描述。在 ATLAS 的左边栏点击 Characterizations，创建一个新的特征描述分析。给该分析命名并点击 按钮保存。

11.7.1 设计

要做特征描述分析，用于本案例定义至少需要一个队列和一个特征用于描述。本案例中我们将使用两个队列。第一组队列定义以开始高血压治疗日期作为索引日期并在之前 1 年内有高血压确诊史。而且我们还要求该队列在开始使用高血压药物后至少有 1 年的持续观察(附录 B.6)。第二组队列除了要求至少进行了 3 年而不是 1 年的持续观察以外，其他要求与第一组相同(附录 B.7)。

队列定义

如第 10 章所述，我们假设已经在 ATLAS 中创建了队列。 点击 Import 并选择如图 11.6 所示的队列。接下来，我们将定义用于描述这两个队列的特征。

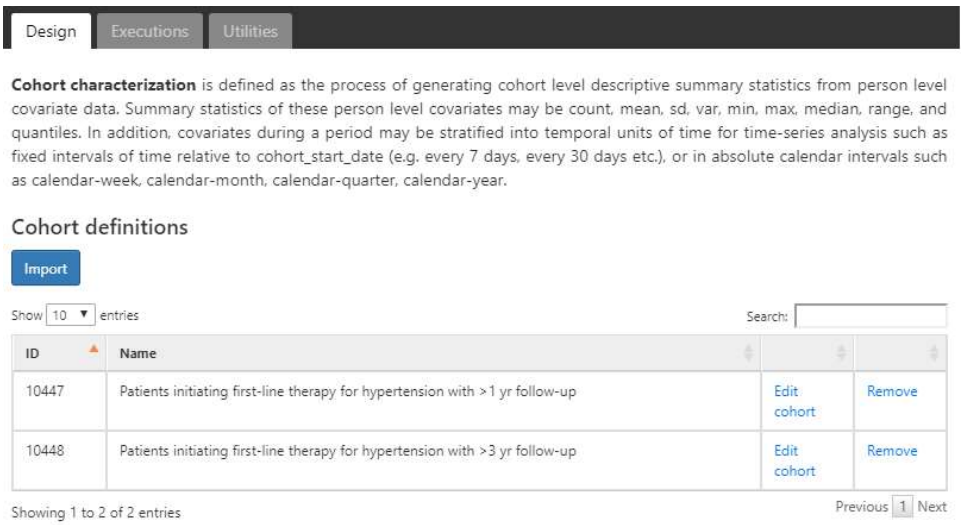


图 11.6: 特征描述设计标签-队列定义选择

特征选择

Atlas 提供了将近 100 个预置的特征分析，用于在 OMOP CDM 模型的临床领域中执行特征描述分析。 每个预设特征分析对选定目标队列的临床观察信息都会执行聚类和概括功能。这些计算可能提供数千个特征来描述队列的基线和索引后特征。在后台功能模块，ATLAS 会利用 “OHDSI Feature

Extraction R”软件包为每个队列执行特征描述功能。我们将在下一节更详细地介绍特征提取和 R 的使用。

点击 Import 选择要描述的特征。下面是我们将用来描述这些队列特性的一系列特征:

Feature analyses

Import

Show 25 entries Search:

ID	Name	Description	Actions
43	Drug Era Short Term	One covariate per drug in the drug_era table overlapping with any part of the short window.	Remove
49	Charlson Index	The Charlson comorbidity index (Romano adaptation) using all conditions prior to the window end.	Remove
67	Condition Occurrence Long Term	One covariate per condition in the condition_occurrence table starting in the long term window.	Remove
71	Demographics Age Group	Age of the subject on the index date (in 5 year age groups)	Remove
72	Demographics Race	Race of the subject.	Remove

(a)

73	Demographics Prior Observation Time	Number of continuous days of observation time preceding the index date.	Remove
74	Demographics Gender	Gender of the subject.	Remove
76	Condition Occurrence Medium Term	One covariate per condition in the condition_occurrence table starting in the medium term window.	Remove
77	Demographics Age	Age of the subject on the index date (in years).	Remove
79	Demographics Time In Cohort	Number of days of observation time during cohort period.	Remove
80	Demographics Index Year	Year of the index date.	Remove
81	Demographics Post Observation Time	Number of continuous days of observation time following the index date.	Remove
87	Procedure Occurrence Any Time Prior	One covariate per procedure in the procedure_occurrence table any time prior to index.	Remove
103	Visit Count Long Term	The number of visits observed in the long term window.	Remove

(b)

图 11.7: 特征描述配置选项卡-特征选择。

上图显示了选定的特征列表，每个特征都有相应关于其如何描述每个队列特征的介绍。以“Demographics”名称开头的特征将计算该队列开始日期时每个人的人口统计学信息。 对于以域名称开头的特征(例如就诊、操作、状况、药物等)，这些将描述该域中所有记录的观察值。每个域的特征在队列开始前有四个时间窗选项，即:

- 之前任何时间：在观察人群的观察期内，早于队列开始时间的所有时间段内
- 长期: 365 天前至队列开始日期当天内。
- 中期: 包括队列开始日期在内的前 180 天内。
- 短期: 包括队列开始日期在内的前 30 天内。

亚组分析

如果我们想研究基于性别差异的特征该如何操做？我们可以使用“亚组分析”模块来定义新的目标亚组并在特征描述分析中使用。

要创建亚组，请单击并添加亚组人员标准。此步骤类似于用于确定队列入组的标准。在本例中我们将定义一组标准来从本队列中识别女性:

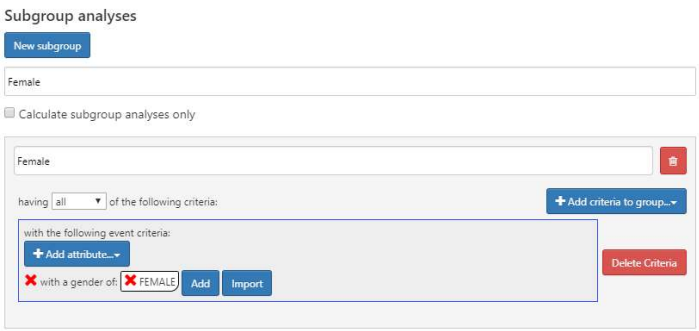


图 11.8: 带有女性亚组分析的特征描述设计。

Atlas 中的亚组分析不同于分层分析。 分层是相互排斥的，而亚组可能依据不同的选择标准而包含同一个人。

11.7.2 执行

一旦我们配置好了特征描述，我们就可以在系统中的一个或多个数据库里执行这个配置。导航到 Executions 选项卡，然后单击 Generate 按钮开始对数据库进行分析:

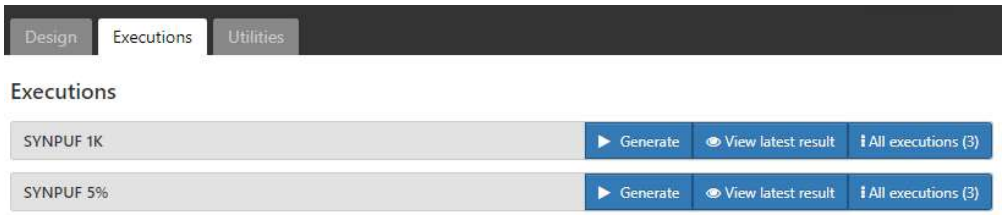


图 11.9: 特征描述设计执行-CDM 源选择。

一旦分析完成，我们可以通过点击“All Executions”按钮查看报告，并从执行列表中选择“View Reports”。或者，可以单击“View latest result”来查看上一次执行情况。

11.7.3 结果

结果提供了配置中每个队列的不同特征的表格视图。在图 11.10 中，表格提供了从队列开始前 365 天内两个队列中出现的所有病情的概要。每个队列的每个协变量都有一个计数和百分比，每个队列也有之前配置的女性亚组。

CONDITION / Condition Occurrence Long Term / stratified by Female

Export | Export comparison | Show 10 entries | Search: cardi

Covariate	Explore	Concept ID	Patients initiating first-line therapy for hypertension with >1 yr follow-up				Patients initiating first-line therapy for hypertension with >3 yr follow-up				Std diff
			Count	Pct	Female		Count	Pct	Female		
					Count	Pct			Count	Pct	
Tachycardia	Explore	444070	17,322	1.04%	9,042	1.18%	6,547	0.78%	3,530	0.90%	-0.0193
Cardiomegaly	Explore	314658	20,958	1.26%	8,007	1.04%	9,016	1.08%	3,465	0.89%	-0.0121
Cardiac arrhythmia	Explore	44784217	30,474	1.83%	13,221	1.72%	14,540	1.74%	6,318	1.62%	-0.0052

Showing 1 to 3 of 3 entries (filtered from 206 total entries) Previous 1 Next

图 11.10: 特征描述结果-长期状况发生

我们使用搜索框来筛选结果来了解有多少比例的人有心律不齐病史，以了解该人群有哪些心血管疾病相关诊断。我们可以使用心律不齐概念旁边的 Explore 链接打开一个新窗口，如图 11.11 所示，其中包含了单个队列概念的更多细节:

Exploring condition_occurrence during day -365 through 0 days relative to index: Cardiac arrhythmia

Cohort: Patients initiating first-line therapy for hypertension with >1 yr follow-up

Export | Show 10 entries | Search:

Relationship type	Distance	Concept name	All stratas		Female	
			Count	Pct	Count	Pct
Explore Ancestor	4	Disorder by body site	32	0.00%	17	0.00%
Explore Ancestor	4	Finding of trunk structure	991	0.06%	605	0.08%
Explore Ancestor	3	Disorder of trunk	23	0.00%	14	0.00%
Explore Ancestor	3	Disorder of thorax	241	0.01%	104	0.01%
Explore Ancestor	3	Disorder of body system	4,135	0.25%	1,992	0.26%
Explore Ancestor	2	Disorder of cardiovascular system	12,979	0.78%	6,073	0.79%
Explore Ancestor	2	Disorder of mediastinum	138	0.01%	62	0.01%
Explore Ancestor	2	Disorder of body cavity	24	0.00%	10	0.00%
Explore Ancestor	1	Heart disease	4,691	0.28%	1,869	0.24%
Explore Selected	0	Cardiac arrhythmia	30,474	1.83%	13,221	1.72%

Showing 1 to 10 of 62 entries Previous 1 2 3 4 5 6 7 Next

图 11.11: 特征描述结果-研究单个概念

因为我们已经描述了队列中所有状况概念的特征，所以 “explore” 选项可以提供所选概念的所有父级和子级概念的视图，如在本例中的心律不齐。这种方式使我们能够纵览和该概念不同层级相关的概念，以观察其他可能出现在高血压人群的心脏疾病。与摘要视图类似，该视图也会显示计数和百分比。

我们也可以使用相同的特征描述检测结果来寻找某些抗高血压治疗禁忌症的情况，如血管性水肿。要做到这一点，我们需按照上面相同的步骤，但搜索的是水肿，如图 11.12 所示:

CONDITION / Condition Occurrence Long Term / stratified by Female

Export | Export comparison | Show 10 entries | Search: edema

Covariate	Explore	Concept ID	Patients initiating first-line therapy for hypertension with >1 yr follow-up				Patients initiating first-line therapy for hypertension with >3 yr follow-up				Std diff
			Count	Pct	Female		Count	Pct	Female		
					Count	Pct			Count	Pct	
Edema	Explore	433595	32,243	1.94%	20,200	2.63%	15,173	1.81%	9,684	2.48%	-0.0066

Showing 1 to 1 of 1 entries (filtered from 206 total entries) | Previous 1 Next

图 11.12: 特征描述结果-禁忌症研究

这次我们依然使用“explorer”功能来了解高血压患者人群的“水肿”的特征，以研究血管性水肿的患病率:

Exploring condition_occurrence during day -365 through 0 days relative to index: Edema

Cohort: Patients initiating first-line therapy for hypertension with >1 yr follow-up

Export | Show 10 entries | Search: ang

Relationship type	Distance	Concept name	All stratas		Female	
			Count	Pct	Count	Pct
Explore Descendant	-2	Angioedema	2,605	0.16%	1,506	0.20%

Showing 1 to 1 of 1 entries (filtered from 56 total entries) | Previous 1 Next

图 11.13: 特征描述结果-禁忌症研究的细节

在这里，我们发现这个人群的一部分在开始服用抗高血压药物前一年内，有血管性水肿的记录。

DEMOGRAPHICS / Demographics Age

Export | Export comparison | Show 10 entries | Search:

Strata	Patients initiating first-line therapy for hypertension with >1 yr follow-up				Patients initiating first-line therapy for hypertension with >3 yr follow-up				Std diff
	Count	Avg	Std Dev	Median	Count	Avg	Std Dev	Median	
Female	768,180	49.39	9.78	51.00	390,693	49.01	9.03	51.00	-0.0291
All stratas	1,661,604	48.96	10.00	50.00	837,459	48.64	9.26	50.00	-0.0232

Showing 1 to 2 of 2 entries | Previous 1 Next

图 11.14: 每个队列和亚组的年龄特征描述结果

虽然“域”的协变量采用的是“有/无”的定性指标(即判断在先前时间段内是否存在对应代码记录)，但有些变量提供了连续值，如队列开始时的人员年龄。在上述例子中，我们用人数、平均年龄、中位年龄和标准差来展示两个队列的年龄情况。

11.7.4 定义定制功能

除了预设功能，ATLAS 可让用户自定义功能。要实现该功能，需要单击特征描述的左侧菜单项，然后单击“Feature Analysis”选项卡，并单击“New Feature Analysis”按钮。提供自定义特征的名称，并点击  按钮。

在本例中，我们将配置一个自定义特征，用于识别在每个队列在队列开始日期后的病史中出现过 ACE 抑制剂疗程记录的人数:

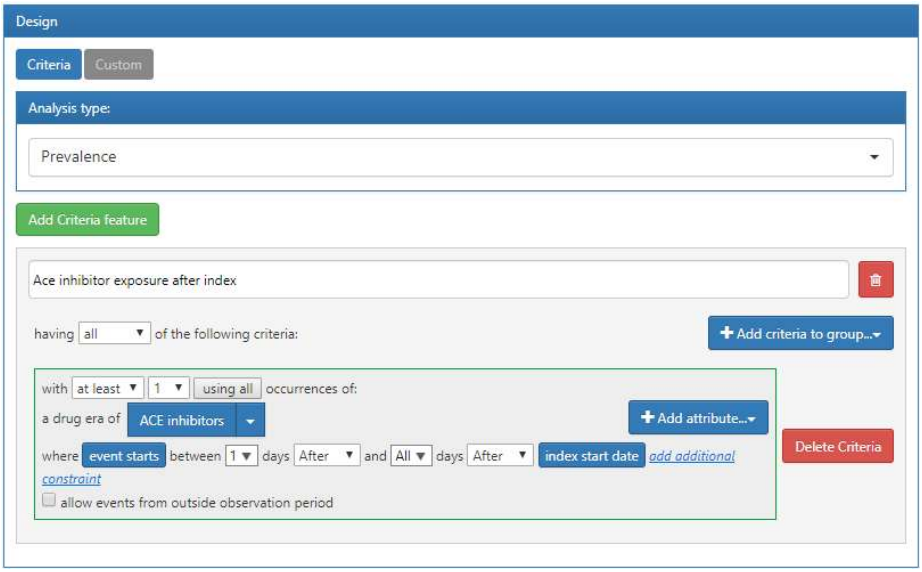


图 11.15: ATLAS 自定义特征。

假设上面配置的标准适用于队列开始日期。如果我们配置了标准并保存，就可以把它应用到之前创建的特征描述配置中。这就需要打开“特征描述配置”页面并回到“Feature Analysis”功能。点击 Import 按钮，从菜单中选择“new custom”功能。它们将出现在特征描述配置的功能列表中。如前所述，我们可以对数据库执行这个配置，为这个自定义特征生成特征描述:

DRUG / Ace inhibitor exposure after index / stratified by Female

Export | Export comparison | Show 10 entries | Search:

Covariate	Explore	Concept ID	Patients initiating first-line therapy for hypertension with >1 yr follow-up				Patients initiating first-line therapy for hypertension with >3 yr follow-up				Std diff
			Count	Pct	Female		Count	Pct	Female		
					Count	Pct			Count	Pct	
Ace inhibitor exposure after index	Explore	0	686,034	41.29%	289,215	17.41%	426,280	50.90%	182,219	21.76%	0.1001

Showing 1 to 1 of 1 entries | Previous 1 Ne

图 11.16: 自定义特征结果展示

11.8 R 软件中队列特征的描述

我们也可以选择使用 R 软件来描述队列特征。在这里我们将描述如何使用 OHDSI R 软件里的特征提取包(Feature Extraction)来生成高血压队列的基线特征(协变量)。特征提取包(Feature Extraction)为用户提供以下三种构造协变量的方式:

- 选择默认的协变量集
- 从一组预先指定的分析中进行选择
- 创建一组自定义分析

特征提取包(Feature Extraction)以两种不同的方式创建协变量：个体水平特征和集合特征。个体水平特征在机器学习中应用较多。在本节中，我们将重点讨论集合特征的使用，这些特征对于描述感兴趣的队列的基线协变量非常有用。另外，我们将重点讨论构造协变量的后两种方法：预先指定分析和自定义分析，并使用预设数据集供读者练习。

11.8.1 队列实例化

我们首先需要实例化队列来进行描述。实例化队列内容在第 10 章中有所介绍。在这个例子中，我们将使用启动高血压一线治疗并随访 1 年的患者(附录 B.6)进行描述。我们在附录 B 中留下了对其他队列的描述以供读者练习。我们假设队列已经在一个名为 scratch.my_cohorts 的表中实例化，定义队列 ID 为 1。

11.8.2 数据提取

我们首先需要告诉 R 软件如何连接到服务器。特征提取(FeatureExtraction)使用 DatabaseConnector 包，它提供了一个名为 createConnectionDetails 的函数。

输入?createConnectionDetails 用于各种数据库管理系统(DBMS)所需的特定设置。例如，可以使用下面的代码连接到 PostgreSQL 数据库：

```
library(FeatureExtraction)
connDetails <- createConnectionDetails(dbms = "postgresql",
                                       server = "localhost/ohdsi",
                                         user = "joe",
                                         password = "supersecret")

cdmDbSchema <- "my_cdm_data" cohortsDbSchema <- "scratch" cohortsDbTable
<-"my_cohorts" cdmVersion <- "5"
```

最后四行定义 cdmDbSchema，cohortsDbSchema，cohortsDbTable 变量,以及 CDM 版本。我们将在稍后使用这些来告诉 R 有关 CDM 格式的数据的路径，感兴趣的队列创建地址，以及使用的 CDM 版本。需要注意的是，对于 Microsoft SQL Server，数据库模式需要同时指定数据库和模式，例如 cdmDbSchema <- "my_cdm_data.dbo".

11.8.3 使用预设分析

createCovariateSettings 函数允许用户从大量预定义的协变量中进行选择。输入?createCovariateSettings 以获得可用选项的概述。例如：

```
settings <- createCovariateSettings (useDemographicsGender = TRUE,
                                     useDemographicsAgeGroup = TRUE,
                                useConditionOccurrenceAnyTimePrior = TRUE)
```

操作会为性别、年龄(5 岁年龄组)以及 condition_occurrence 表中在队列开始日期之前(包括该时间)的每个概念创建二元协变量。

许多预先设定的分析是指短期、中期或长期的时间窗。一般默认这些窗口的定义如下：

- **长期**：包括队列开始日期在内的 365 天前。
- **中期**：包括队列开始日期在内的 180 天前。
- **短期**：包括队列开始日期在内的 30 天前。
- 但是用户可以根据情况进行更改。比如：

```
settings<-createCovariateSettings(useConditionEraLongTerm=TRUE,
                                   useConditionEraShortTerm = TRUE,
                                    useDrugEraLongTerm = TRUE,
                                   useDrugEraShortTerm = TRUE,
                                      longTermStartDays =-180,
                                       shortTermStartDays =-14,
                                                  endDays = -1)
```

本例将长期窗口重新定义为队列开始日期的 180 天前(但不包括开始日期)，将短期窗口重新定义为队列开始日期的 14 天前(但不包括开始日期)。

同样我们也可以指定某些概念是否应该用于构建协变量：

```
settings <- createCovariateSettings(useConditionEraLongTerm = TRUE,
                                     useConditionEraShortTerm =
                                     TRUE, useDrugEraLongTerm =
                                     TRUE, useDrugEraShortTerm =
                                     TRUE, longTermStartDays = -
                                     180,
                                     shortTermStartDays =-14,
                                     endDays = -1,
                                     excludedCovariateConceptIds =
                                     1124300, addDescendantsToExclude =
                                     TRUE, aggregated = TRUE)
```

11.8.4 创建聚合协变量

上面所有示例中 aggregated = TRUE 的使用表明特征提取可以提供简要的统计信息。排除此标志将会为队列中每个个体计算协变量。

下面的代码块将生成一个队列的聚合统计数据。

```
covariateSettings <- createDefaultCovariateSettings()

               covariateData2 <- getDbCovariateData (connectionDetails =
                                                       connectionDetails,
                                     cdmDatabaseSchema = cdmDatabaseSchema,
                                  cohortDatabaseSchema = resultsDatabaseSchema,
                                      cohortTable = "cohorts_of_interest",
                                                          cohortId = 1,
                                       covariateSettings = covariateSettings,
                                                          aggregated = TRUE)
 summary(covariateData2)
```

输出结果与下面内容类似：

```
## CovariateData Object Summary
##
## Number of Covariates: 41330
## Number of Non-Zero Covariate Values: 41330
```

11.8.5 输出格式

聚合协变量数据对象的两个主要组成为 covariates 和 covariatesContinuous，分别对应二分类和连续性协变量:

```
covariateData2$covariates
covariateData2$covariatesContinuous
```

11.8.6 自定义协变量

特征提取(FeatureExtraction)还提供了自定义协变量和使用自定义协变量的能力。这些细节是一个更高级的话题，在用户文档中有所涉及: http://ohdsi.github.io/ FeatureExtraction/.

11.9 ATLAS 中的队列路径

路径分析的目的是了解一个或多个感兴趣队列的治疗顺序。其应用的方法是基于2016年Hripcsak等人所提出的设计。在 ATLAS 中，这些方法被扩展并编码为一个特征，称之为队列路径。

队列路径旨在提供分析能力，以汇总一个或多个目标队列的队列开始日期之后的事件。为了做到这一点，我们创建了一系列队列来识别目标人群中感兴趣的临床事件，称之为事件队列。聚焦于目标队列个体的寻找方法。

图 11.17: 个人背景下的路径分析。

在图 11.17 中，这个人是具有开始和结束日期的目标队列的一部分。随后，编号的线段代表示一段时间内这个人在事件队列中被识别的地方。事件队列允许我们描述 CDM 中所代表的任何感兴趣的临床事件，这样我们就不会局限于为一个单独领域或概念创建路径。

首先，点击 ATLAS 左侧栏 

创建一个新的队列路径研究。提供一个描述性的名称并按下保存按钮。

11.9.1 设计

首先，我们将继续使用启动高血压一线治疗的队列，随访 1-3 年(Appendix B.6, B.7)。使用这个按钮导入 2 个队列。

接着我们通过为每个感兴趣的一线高血压药物创建队列来定义事件队列。为此，我们将首先创建一个 ACE 抑制剂使用者队列，并将队列结束日期定义为持续暴露的结束日期。我们将对其它 8 种高血压药物做同样的处理，这些定义见 Appendix B.8-B.16。完成后，使用 

按钮导入到路径设计的事件队列部分。

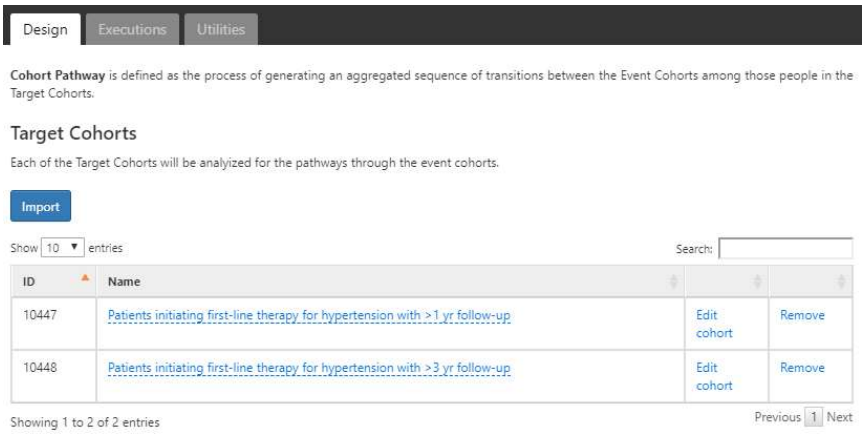


图 11.18: 选定目标人群的路径分析。

当完成时，你的设计应该如上面那样。接下来我们需要确定一些额外的分析设置：

- **组合窗口**：该设置允许你定义一个时间窗，以天为单位，其中事件之间的重叠被认为是事件组合。例如，如果 2 个事件队列(事件队列 1 和 2)代表的 2 种药物在组合窗口内重叠，路径算法则会把它们合并为“事件队列 1+事件队列 2”

- **最小单元计数**：少于这个人数的事件队列会在输出中进行审查(删除)以保护隐私。

最大路径长度：这是指分析时要考虑的连续事件的最大数量。

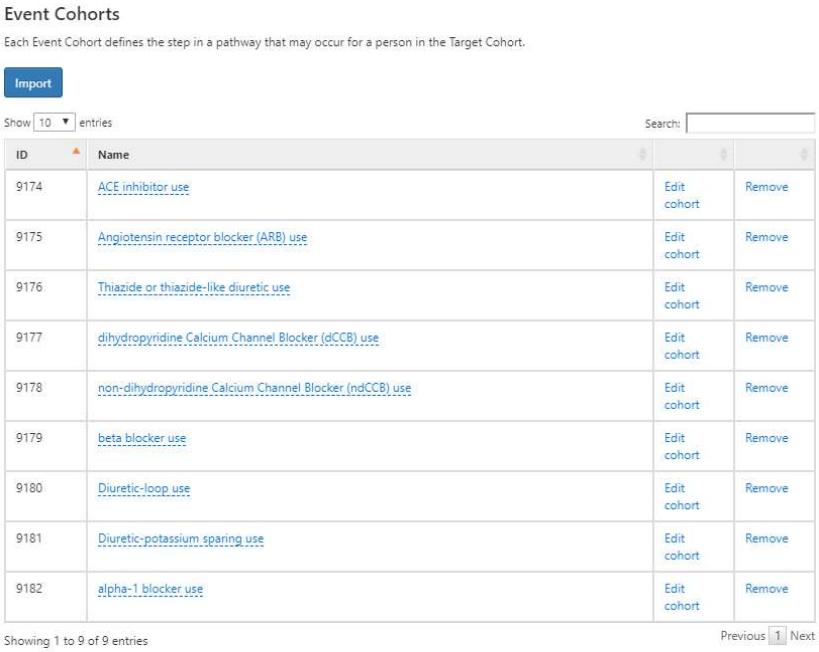


图 11.19: 启动高血压一线治疗事件队列的路径分析

11.9.2 执行

一旦我们设计好路径分析，我们就可以针对数据环境中的一个或多个数据库执行此设计。这与我们在 ATLAS 中队列特征描述的工作方式相同。当完成时，我们可以查看分析的结果。

11.9.3 结果查看

路径分析的结果分为以下 3 个部分：图例部分显示目标队列总人数以及在路径分析中发生一个或多个事件的人数。摘要下方是中央部分旭日图中出现的每个队列的颜色标识。

“旭日”(sunburst)图是一种可视化图形，代表个体随时间推移出现的各种事件路径。图形中心代表队列入口，第一个彩色编码环显示每个事件队列的人员比例。在我们的例子中，圆心代表启动高血压一线治疗的人群。然后，旭日图的第一个环显示按照事件队列(即 ACE 抑制剂，血管紧张素受体阻滞剂等)定义的开始一线治疗的人群比例。第二个环代表第二个事件队列的人群。在某些特定的事件顺序中，在数据中可能观察不到同一个人的第二个事件队列，这部分人群由环中的灰色部分表示。

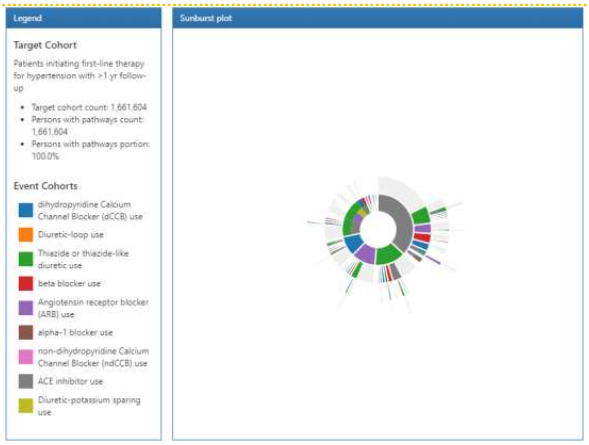


图 11.20: 路径分析结果说明和可视化旭日图

图 11.21: 显示显示路径细节的路径分析结果

点击旭日图的某一部分将会在右侧显示路径细节。在这个例子中我们可以看到，在我们的目标队列中，使用 ACE 抑制剂开始一线治疗的人群所占比例最大，而同样在此组中，使用噻嗪类利尿剂的比例更低。

11.10 ATLAS 中的发病率分析

在发病率的计算中，我们描述：目标队列中，在风险期内经历结局队列的人群。接下来我们会设计一个发病率分析，来描述 ACE 抑制剂(ACEi)和噻嗪类利尿剂(THZ)新使用者中血管性水肿和急性心肌梗死的预后。我们将在患者暴露于这种药物的 TAR 中评估这些结局。此外，我们还将增加一个暴露于血管紧张素受体阻滞剂(ARBs)的结局，用于评估暴露于目标队列(使用 ACEi 和 THZ)期间新使用 ARBs 的发病率。这一结局的定义有助于理解 ARBs 在目标人群中的使用情况。

首先，点击 ATLAS 左栏的 Incidence Rates 以创建一个新的发病率分析。提供一个描述名称并按下保存按钮 。

11.10.1 设计

我们假设本例中使用的队列已经如第 10 章所述在 ATLAS 中创建。附录提供了目标队列(Appendix B.2, B.5)和结局队列(Appendix B.4, B.3, B.9)的完整定义。

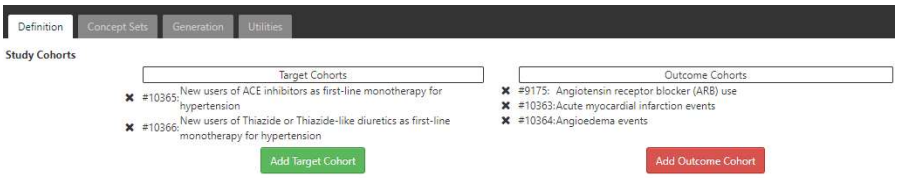


图 11.22: 目标队列和结局队列发病率定义

定位到定义标签，点击选择 New users of ACE inhibitor 队列和 New users of Thiazide or Thiazide-like diuretics 队列。关闭这个对话框并查看这些队列是否已添加入设计。接下来我们通过点击对话框来添加结局队列，选择 acute myocardial infarction events, angioedema events 和 Angiotensin receptor blocker (ARB) use 结局队列。同样，关闭对话框并查看这些队列是否已添加入设计中的结局队列部分。

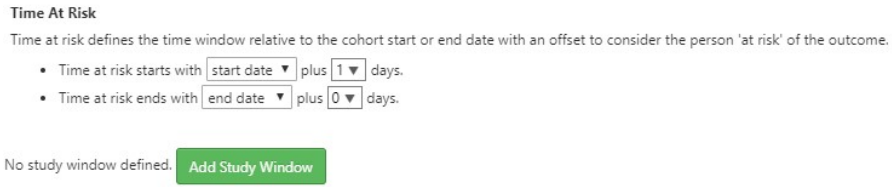


图 11.23: 目标队列和结局队列发病率定义

接下来，我们将定义用于分析的风险时间窗。如上所示，风险时间窗是相对于队列开始和结束日期而定义的。这里我们定义风险开始时间为我们的目标队列开始后 1 天。接着我们定义风险结束时间为队列结束时间。在当前的案例中，当药物暴露结束时，定义的 ACEi 和 THZ 队列有一个队列结束日期。

ATLAS 还提供了对目标队列进行分层的方法，作为分析参数的一部分：

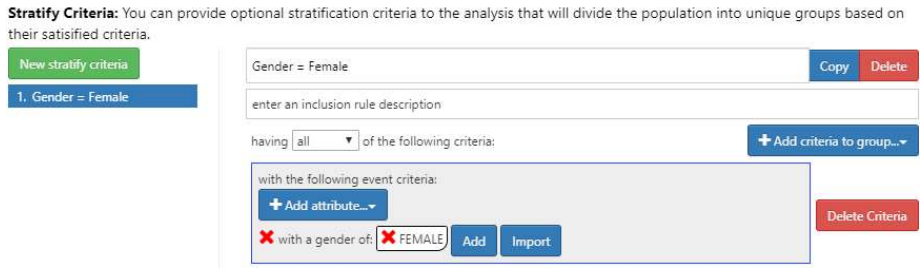


图 11.24: 性别分层下女性发病率的定义

要实现这一点，点击 New Stratify Criteria 按钮并按照章节 11 中所描述的相同步骤操作。现在我们已经完成了这个设计，可以转向一个或多个数据库执行设计。

11.10.2 执行

点击 Generation 选项卡，然后点击 ▶ Generate 按钮显示用于执行分析的数据库列表：

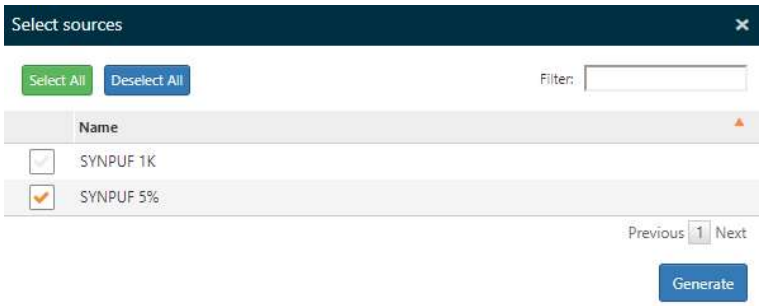


图 11.25: 发病率分析执行。

选择一个或多个数据库并点击 Generate 按钮开始分析，分析设计中指定的目标和结局的所有组合。

11.10.3 结果查看

在 Generation 选项卡上，查看结果时屏幕顶端允许你选择一个目标和结局。下方是对分析中所使用每个数据库的发病率总结。

从各自的下拉列表中选择目标队列中的 ACEi users 和 Acute Myocardial Infarction (AMI)。点击 Reports 按钮显示发病率的分析结果：

	Persons	Cases	Proportion [+|-] per 1k persons	Time At Risk (years)	Rate [+|-] per 1k years
Summary Statistics:	816,867	1,703	2.08	623,402	2.73

	Stratify Rule	N	Cases	Proportion [+|-] per 1k persons	Time At Risk (years)	Rate [+|-] per 1k years
1.	Gender = Female	313,518	416	1.33	230,157	1.81

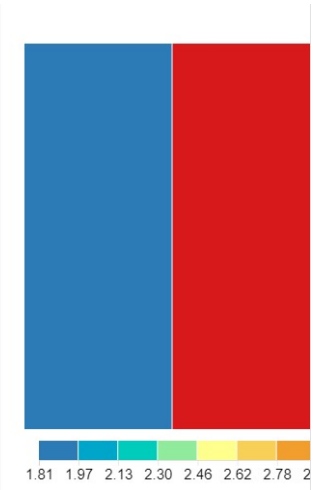


图 11.26: 发病率分析结果输出–有 AMI 结局的 ACEi 新使用者

数据库汇总显示 TAR 期间所观察到的队列中总人数和病例总数。比例以每 1000 人中的例数显示。目标队列中风险时间的计算以年为单位。发病率以每 1000 人年的例数表示。

我们也可以查看设计中定义的分层发病率指标。对每一层计算上面所提到的同样的指标。此外，可视化的树型图提供了方框区域代表的每个分层的比例。颜色代表发病率，如底部刻度所示。

我们可以收集同样的信息来观察 ACEi 使用人群中新使用 ARBs 的发病率。使用顶部的下拉菜单，将结局改为 ARBs 的使用并点击 Reports 按钮来显示细节。

	Persons	Cases	Proportion [+|-] per 1k persons	Time At Risk (years)	Rate [+|-] per 1k years
Summary Statistics:	818,520	43,868	53.59	621,094	70.63

Stratify Rule	N	Cases	Proportion [+|-] per 1k persons	Time At Risk (years)	Rate [+|-] per 1k years
1. Gender = Female	313,832	21,049	67.07	228,369	92.17

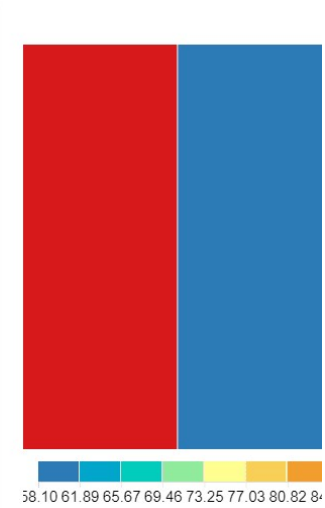


图 11.27: 发病率—— ACEi 暴露期间接受 ARB 治疗的新用户。

如图所示，计算的指标相同但解释不同，因为输入(ARB use)引用了药物使用的估计值而不是健康结局。

11.11 总结

OHDSI 为整个数据库或感兴趣的队列提供描述工具。

队列特征描述索引日期之前(基线)和索引日期之后(索引后)的感兴趣队列。

ATLAS 的特征描述模块和 OHDSI 的方法库提供为多个时间窗计算基线特征的能力。

ATLAS 的路径和发病率模块为索引后期提供描述性统计。

11.12 练习

必备条件

对于这些练习需要访问 ATLAS 实例。你可以使用 http://atlas-demo.ohdsi.org 所提供的实例或其它任何你所能获得的实例。

练习 11.1.我们想要了解塞来昔布在真实世界中的使用情况。首先我们想要了解数据库中关于这个药物的数据。使用 ATLAS 数据源模块来寻找关于塞来昔布的信息。

练习 11.2.我们想要更好地了解塞来昔布使用者的疾病自然史。创建一个基于 365 天洗脱期(详见第 10 章)的塞来昔布新使用者的简单队列，并用 ATLAS 创建这个队列的特征描述，显示队列中共病情况和药物暴露情况。

练习 11.3.我们对人们开始使用塞来昔布后任何时间发生的胃肠出血的频率感兴趣。创建一个胃肠出血事件的队列，简单定义为任意一次胃肠出血(概念 192671)的发生或任何它的衍生概念的发生。使用前面练习中定义的暴露队列来计算塞来昔布使用后胃肠出血事件的发病率。

参考答案见附录 E.7。

参考文献

1. Elm, Erik von, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. 2008. "The Strengthening the Reporting of Observational Studies in Epidemiology (Strobe) Statement: Guidelines for Reporting Observational Studies." Journal of Clinical Epidemiology 61 (4): 344–49. https://doi.org/10.1016/j.jclinepi.2007.11.008.
2. Hripcsak, George, Patrick B. Ryan, Jon D. Duke, Nigam H. Shah, Rae Woong Park, Vojtech Huser, Marc A. Suchard, et al. 2016. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113 (27): 7329–36. https://doi.org/10.1073/pnas.1510502113.
3. Who, A. 2013. "Global Brief on Hypertension." World Health Organization. https://www.who.int/cardiovascular_diseases/publications/global_brief_hypertension/en/.

第十二章　人群水平评估

章节负责人：马丁•舒米（Martijn Schuemie），大卫•麦迪根（David Madigan），马克•苏卡德（Marc Suchard）和帕特里克•瑞安（Patrick Ryan）

观察性健康数据，例如医保和电子病历记录，为生成关于治疗效果是否有意义地改善患者生活的真实世界证据提供了机会。 在本章中，我们着重于人群水平的效果评估，这是指暴露因素（如药物或操作的医疗干预措施）对特定医疗结局的因果效应。接下来，我们考虑两种不同的评估内容：

直接效应估计:与非暴露组相比，估计暴露因素对结局发生风险的影响。

比较效应估计:与某种暴露(对照暴露)相比，估计另一种暴露（目标暴露）对结局发生风险的影响。

在这两种情况下，患者个体水平的因果效应估计，是通过对比事实结局和反事实结局来实现，即比较暴露组患者的结局效应和假设患者没有暴露或受到另一种暴露时的潜在结局效应的差异。由于在实际情况中任何一个患者只有一种事实结局（因果推论的基本难题），所以我们通过设计不同的效应估计方法来推断反事实结局。

人群水平效应评估的用例包括治疗选择、安全性监测和疗效对比。方法可以是一次测试、一个特定的假设（如“信号评估”）或同时探索多个假设（如“信号检测”）。在所有情况下，目标都是一样的：对因果效应进行高质量的评估。

12.1 队列研究设计

在本章中，我们首先描述了各种人群水平估计的研究设计，所有这些设计都可用 OHDSI 方法库中的 R 语言包实现。随后，我们详细介绍了估计研究的设计，并逐步指导如何使用 ATLAS 和 R 来实现。最后，我们回顾了研究的各种产出，包括研究诊断程序和效果评估。

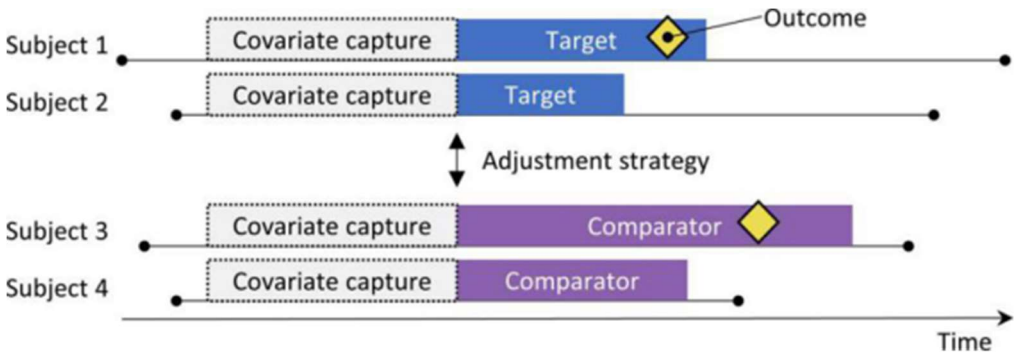


图 12.1：新用户队列设计。从实施治疗干预措施的那一刻起开始观察，比较目标治疗组和对照治疗组的结局。

为了校正两个治疗组之间的差异，可以使用几种校正策略，如分层、匹配、倾向性评分加权或将基线特征添加到结局模型中。倾向性模型或结局模型中包含的特征须在治疗开始前捕获。

新用户队列研究设计与临床随机对照试验相似。(Hernan and Robins, 2016)

从受试者开始受到治疗干预措施时观察，并随访特定的时间，比较目标治疗组和对照治疗组的差异，如受试者接受治疗的时间。我们可以通过表 12.1 中列出的五个要素来明确我们希望在队列研究中回答的问题。

表 12.1:队列研究设计中的主要设计选择。

选择	描述
目标队列	接受目标治疗的队列
对照队列	接受对照治疗的队列
结局队列	代表目标结局的队列
风险暴露期间	在什么时间段（通常与目标和对照队列开始和结束随访的时间相关）考虑结局发生的风险
模型	使用模型校正目标组和对照组之间的差异后来评估效果

依据结局类型指定模型的选择。例如，我们可以使用 Logistic 回归来评估结局是否已经发生，并产生一个比值比（odds ratio）。Logistic 回归假设风险时间段在目标和对照组之间是相同的长度，或是不相关的。或者，我们可以在假设发病率为常数的情况下选择 Poisson 回归来估计发病率比率（rate ratio）。在假设目标与对照组之间风险比例不变的情况下，使用结局首次发生的时间来估计危险比（hazard ratio）通常使用 Cox 回归。

新用户队列犯法本质上是对两种不同治疗方法的效果评估。用这种方法来比较治疗和不接受治疗的差异是很困难的，因为很难找到一组未暴露的人与暴露组是可比的。如果想要使用这种设计来进行直接效果评估，首选的方法是选择一个与接收目标暴露的受试者之争相同的对照治疗，其中对照治疗被认为对结局没有影响。遗憾的是，并不是总能找到这样的对照

一个关键的问题是，接受目标治疗的患者可能与接受对照治疗的患者存在系统性差异。例如，假设目标队列的平均年龄为 60 岁，而对照队列的平均年龄为 40 岁。在任何与年龄相关的医疗结局（如卒中）方面，可能会在队列之间比较时显示出显著的差异。一个不知情的研究者可能会得出结论：与对照组相比，目标治疗与卒中之间存在因果关系。更常见的是，研究者可能会得出这样的结论:发生卒中的目标队列患者，如果他们接受了对照治疗，可能就不会卒中。这个结论很可能是完全错误的！也许那些目标队列患者不成比例地发生卒中仅仅是因为他们的年龄更大；也许那些发生卒中的目标队列患者，即使他们接受了对照治疗，也依然会发生。在这种情况下，年龄是一个“混杂因素”。在观察性研究中，处理混杂因素的一种方法是倾向性评分。

12.1.1 倾向性评分

在随机试验中，可以通过抛硬币（虚拟地）将患者分配到不同的组。这样的设计使得患者接受目标治疗的可能性（与接受对照治疗相比）与患者的年龄等特征没有任何关系。硬币是不能区分患者的，更重要的是，我们确切地知道患者接受目标治疗的准确概率。因此，随着试验中患者人数的增加，置信度也在不断提高，这两组患者理论上不会在任何患者特征上有系统性差异。这保证了试验中可测量的特征（如年龄）以及试验未能测量的特征（如患者遗传因素）均获得了平衡。

对于一个给定的患者，倾向性评分（propensity score，PS）是患者相较接受对照治疗而言，接受目标治疗的概率。（Rosenbaum 和 Rubin, 1983 年）在一个平衡的双臂随机试验中，每位患者的倾向性评分均为 0.5。在倾向性评分校正的观察性研究中，我们基于在治疗开始时和开始前观察到的数据来评估患者接受目标治疗的概率（与他们实际接受的治疗无关）。这是一个直接的预测建模的应用；我们拟合一个模型（如 Logistic 回归）来预测一个受试者是否接受了目标治疗，并使用该模型为每个受试者生成预测概率（predicted probabilities, the PS）。与标准随机试验不同，不同的患者接受目标治疗的概率不同。PS 可以用在几个方面，包括匹配 PS 相近的目标受试者与对照受试者，基于 PS 对研究人群进行分层，或使用源于 PS 的逆处理概率加权（Inverse Probability of Treatment Weighting，IPTW）来给受试者加权重。匹配是我们可以为每位目标受试者选择一位对照，或我们可以为每位受试者匹配多位对照受试者，而这个方法被称之为可变比例（variable-ratio）匹配。（Rassen et al., 2012）

例如，假设我们使用一对一的 PS 匹配，Jan 接受目标治疗的先验概率为 0.4，且确实接受了目标治疗。如果我们能找到一个患者 Jun，他接受目标治疗的先验概率也是 0.4，但实际上他接受了对照治疗，那么 Jan 和 Jun 的结局的比较就像一个迷你随机试验，至少是在平衡了已测量混杂因素的情况下。这个比较将产生与随机化效果一样好的一个 Jan-Jun 因果对比的估计。估计过程如下：对于每位接受目标治疗的患者，匹配一位或多位接受对照治疗但具有与之相同先验概率的对照患者。在每个匹配组中比较目标患者与对照患者的结局。

倾向性评分控制了可测量的混杂因素。事实上，如果已知测量的特征对治疗分配是“可忽略不计”的，倾向性评分将产生一个对因果效应不偏倚的估计。“可忽略不计”实质上是指不存在未测量的混杂因素，并且对测量的混杂因素进行了适当的校正。不幸的是，这是一个不可测试的假设。关于这个问题的进一步讨论见第 18 章。

12.1.2 变量的选择

在过去，PS 是基于手动选择的特征进行计算的，虽然 OHDSI 工具可以支持这种做法，但我们更喜欢使用许多通用特征（即，这些特征不是基于研究中特定的暴露和结局而选择的）。（Tian et al., 2018）这些特征包括人口统计学特征，以及在开始治疗之前和当天观察到的所有的诊断、药物暴露、检测和医疗操作。一个模型通常包含 10,000 到 100,000 个唯一特征，我们使用在 Cyclops 包中实现的大规模正则化回归（Suchard et al., 2013）来拟合这些特征。本质上，我们让数据决定哪些特征具有治疗分配的预测性，并应该囊括在模型中。

一些人认为，不依赖临床专业知识来确定“正确”因果结构的数据驱动的协变量选择方法存在错误地引入所谓工具变量和碰撞变量的风险，从而增加了变异和引入偏倚的潜在风险。(Hernan et al., 2002)然而，这些担心不太适用于真实世界场景。（Schneeweiss, 2018）此外，在医学领域，真正的因果结构很少被人所知，当不同的研究人员被要求为一个特定的研究问题确定“正确的”协变量时，每个研究人员都会给出一个不同的列表，从而使这个过程不可重复。最重要的是，我们的诊断检查方法，如检查倾向性模型，评估所有协变量的平衡性，和纳入阴性对照，将能定位大部分碰撞变量和工具变量相关问题。

我们通常在协变量捕获窗口中囊括开始治疗的当天，因为许多相关的数据信息，如导致治疗发生的诊断，都记录在哪个日期。在这一天，目标和对照治疗本身也被记录，但这些不应该包括在倾向性模型中。因为它们正是我们试图预测的东西。因此，我们必须明确地将目标和对照

12.1.3 卡尺

由于倾向性评分的分数从 0 到 1 连续，精确地匹配几乎是不可能的。所以，使用倾向性评分匹配目标患者时会有一些容忍度，称为“卡尺”。继 Austin(2011)之后，我们在 logit 量表上使用 0.2 个标准差作为默认卡尺。

12.1.4 重叠:偏好评分

倾向性评分法要求存在可匹配的患者！因此，一个关键的检查为显示两组的倾向性评分分布。为了便于解释，OHDSI 工具绘制了一个称为“偏好评分”的倾向性评分转换图。（Walker et al., 2013）偏好评分根据两种治疗的“市场份额”进行调整。例如，如果 10%的患者接受了目标治疗（90%接受了对照治疗），那么偏好评分为 0.5 的患者接受目标治疗的概率为 10%。偏好评分的数学公式如下：

$$ln\frac{(F)}{(1-F)} = ln\frac{(S)}{(1-S)} - ln\frac{(P)}{(1-P)}$$

其中 F 为偏好评分，S 为倾向性评分，P 为接受目标治疗的患者比例。

Walker et al. (2013) 讨论了“经验均衡”的概念。如果至少有一半的暴露对象是偏好评分在 0.3 至 0.7 之间的患者，那么他们接受暴露配对（exposure pairs）符合经验均衡的显现。

12.1.5 均衡

总是检查 PS 校正效果的良好习惯有助于构建均衡的患者组。图 12.19 显示了用于检查均衡性的标准 OHDSI 输出。对于每一项患者特征，这绘制了 PS 校正前后两个暴露组之间的均值标准化差异。一些指南建议校正后的标准化差异上限为 0.1。（Rubin, 2001）

12.2 自身对照队列设计

图 12.2：自身对照队列设计。将目标暴露期间的结局发生率与目标暴露之前的结局发生率进行比较。

自身对照队列（self-controlled cohort，SCC）设计（Ryan et al., 2013a）比较暴露期间的结局发生率与暴露之前的结局发生率。表 12.2 所示的四个选项定义了一个自身对照队列的问题。

表 12.2:自身对照队列设计的主要设计选择

选择	描述
目标队列	接受目标治疗的队列
结局队列	代表目标结局的队列
风险暴露期间	在什么时间段（通常与目标和对照队列开始和结束随访的时间相关）考虑结局发生的风险
对照时间段	用作对照的时间段

由于构成暴露组的受试者也用于对照组，因此不需要对人与人之间的差异进行校正。但是，该方法却易受其他差异的影响，例如不同时间段之间结局的基线风险的差异。

12.3 病例对照设计

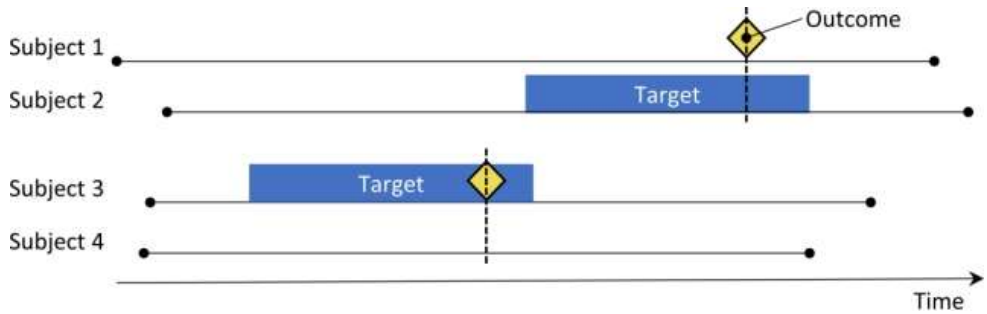


图 12.3：病例对照设计。有结局的受试者（“病例”）与没有结局的受试者（“对照”）在暴露状态方面进行比较。通常

情况下，病例和对照组在年龄和性别等不同特征上进行匹配。

病例对照研究（Vandenbroucke and Pearce, 2012）考虑的问题是“有特定疾病结局的人是否比没有的人更频繁地暴露于一个特定因素？”因此，中心思想是比较“病例”与“对照”，即比较发生了目标结局的受试者与没有发生目标结局的受试者。表 12.3 中的选项定义了一个病例对照问题。

表 12.3：病例对照设计中的主要设计选择

选择	描述
结局队列	代表目标结局的队列
对照队列	代表对照。通常的对照队列是使用一些选择逻辑自动从结局队列衍生
目标队列	代表治疗的队列
巢式队列	可选的，从队列中选取亚组人群作为病例和对照
风险暴露期间	在什么时间段（通常与索引日期有关）考虑暴露状态

通常，人们会控制年龄和性别等特征来为病例匹配对照，以使他们更具可比性。另一种普遍的做法是纳入一个特定亚组人群来进行分析，例如，所有人都被诊断出有某个目标暴露的指征的亚组。

12.4 病例交叉设计

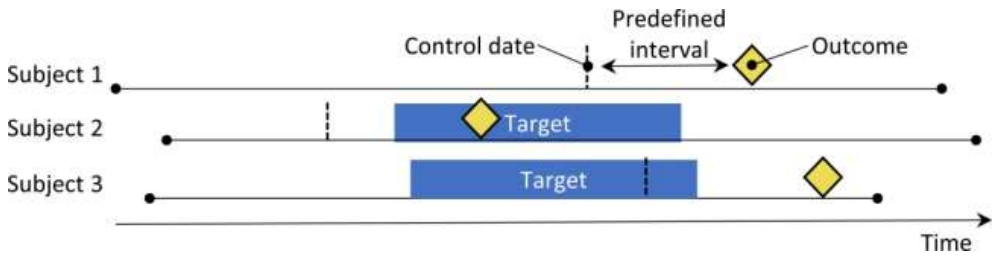


图 12.4：病例交叉设计。将结局时间与预先设定的早于结局日期的对照日期进行比较。

病例交叉（Maclure, 1991）设计评估结局发生时的暴露率是否不同于结局发生前某个预定天数的暴露率。它试图确定结局发生的那段时间是否有什么特别。表 12.4 显示了定义病例交叉问题的选项。

表 12.4:病例交叉设计的主要设计选择。

选择	描述
结局队列	代表目标结局的队列
目标队列	代表治疗的队列
风险暴露期间	在什么时间段（通常与索引日期有关）考虑暴露状态
对照时间段	用作对照时间的时间段

病例本身同时也作为他们自己的对照。作为自身对照的设计，由于人与人之间的差异，更易于混淆。但有一个问题是，因为结局日期总是晚于对照日期，如果暴露的总频率随时间增加，该方法将产生正偏倚；如果暴露的总频率随时间减少，则产生负偏倚。为了解决这个问题，开发了 case-time-control 设计（Suissa, 1995），它将匹配年龄和性别等的控制添加到病例交叉设计中，以校正暴露趋势的影响。

12.5 自身病例对照系列设计

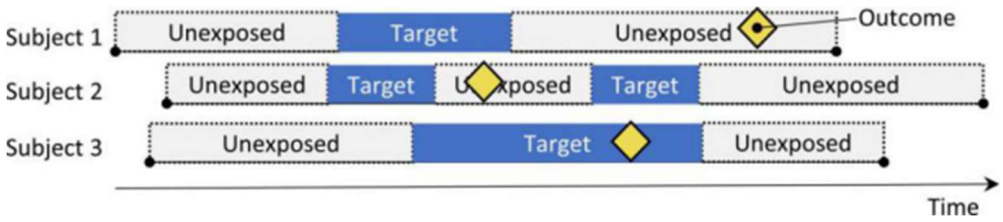


图 12.5：自身病例对照系列设计。暴露期间的结局发生率与未暴露时的结局发生率进行比较。

自身病例对照系列（Self-Controlled Case Series，SCCS）设计（Farrington, 1995; Whitaker et al., 2006）比较暴露期间的结局发生率与所有未暴露期间的结局发生率，包括暴露前、暴露期间和暴露后。这是对个体患者的一种 Poisson 回归。因此，它试图回答这样一个问题：“假定一位患者已经产生了结局，那该结局在暴露期间较非暴露期间更有可能发生吗?”表 12.5 中的选项定义了一个 SCCS 问题。

与其他自身病例对照设计一样，SCCS 对由于人与人之间的差异而产生的混杂，但易受到由于时变效应而产生的混杂的影响。试图解决这些问题，例如纳入年龄和季节进行一些校正是可能的。SCCS 的一个特殊变体不仅包括目标暴露，还包括数据库中记录的所有其他药物暴露（Simpson et al., 2013），这可能会给模型增加数千个额外变量。利用交叉验证选择的正则化超参数的 L1 -正则化方法可应用于除目标暴露之外的所有暴露的系数。

SCCS 的一个重要假设是，观察期的结束与结局的日期无关。对于某些结局，特别是那些可致命的结局，如卒中，这种假设可能被打破。一个 SCCS 扩展方法被开发出来纠正此类依赖关系。(Farrington 等，2011)

表 12.5：自身病例对照系列设计的主要设计选择。

选择	描述
目标队列	代表治疗的队列
结局队列	代表目标结局的队列
风险暴露期间	在什么时间段（通常与目标队列的开始和结束日期相关)我们考虑结局的风险?
模型	对效果进行估计，包括对时变性混杂因素的任何校正

12.6 设计一项高血压研究

12.6.1 问题定义

血管紧张素转换酶抑制剂（ACEi）广泛用于高血压或缺血性心脏病患者的治疗，特别是那些存在合并症的患者，如合并充血性心力衰竭，糖尿病或慢性肾脏疾病。(Zaman et al.，2002）血管性水肿是一种严重且有时危及生命的不良事件，通常表现为嘴唇、舌头、口腔、喉头、咽或眶周区域的肿胀，与这些药物的使用有关。(Sabroe and Black, 1997）然而，关于药物相关血管性水肿的绝对和相对风险的信息是有限的。现有证据主要是基于对特定群体的调查（例如，主要为男性的退伍军人或医疗补助受益人)，基于该类患者的发现可能不能推广到其他人群，或者，基于事件发生很少的调查，这些调查提供的风险估计并不稳定。(Powers et al.，2012）几项观察性研究比较了 ACEi 和β-受体阻滞剂对血管性水肿发生风险的影响（Magid et al., 2010; Toh et al., 2012)，但β-受体阻滞剂已不再被认为是高血压的一线治疗药物。(Whelton et al.， 2018）一种可行的替代治疗是噻嗪类药物或类噻嗪类利尿剂（THZ)，它们在管理高血压及其相关风险方面，如急性心肌梗死（AMI)，可以不增加血管性水肿的风险而同样有效。

下面将演示如何将我们的人群水平估计框架应用于观察性医疗保健数据，以解决以下对比估计问题:

与噻嗪类利尿剂和类噻嗪类利尿剂的新服用者相比，ACE 抑制剂的新服用者发生血管性水肿的风险如何?

与噻嗪类利尿剂和类噻嗪类利尿剂的新服用者相比，ACE 抑制剂的新服用者发生急性心肌梗死的风险如何?

由于这些是比较效应估计问题，我们将采用第 12.1 节中描述的队列方法。

12.6.2 目标和对照

如果患者首次观察到的高血压治疗是使用 ACEi 或 THZ 类中任一活性成分的单方治疗，将被考虑为新使用者。我们将单方治疗定义为在开始治疗后 7 天内未开始服用任何其他降压药。我们要求患者首次暴露前在数据库中至少有一年的连续观察记录，和治疗开始或开始前一年的至少一次高血压的诊断记录。

12.6.3 结局

我们将血管性水肿定义为在住院或急诊室（ER）就诊期间出现的任何血管性水肿情况，并要求在此七天前无血管性水肿的诊断记录。我们将急性心肌梗死定义为住院或急诊室期间出现的任何急性心肌梗死情况，并要求在此之前 180 天内无急性心肌梗死的诊断记录。

12.6.4 风险时间段

我们将风险时间段定义为在开始治疗后的当天开始，在暴露停止时停止，容许在药物暴露期间至少有 30 天的间隔。

12.6.5 模型

我们使用默认的协变量集来拟合 PS 模型，包括人口统计学指标、病情、药物、操作、检查、观察和几个合并症共同发病率评分。我们从协变量中排除了 ACEi 和 THZ。我们执行可变比例匹配，并在匹配的数据集上设置 Cox 回归条件。

12.6.6 研究总结

表 12.6:我们的比较队列研究的主要设计选择。

选择	描述
目标队列	将 ACE 转换酶抑制剂作为高血压的一线单方治疗的新服用者
对照队列	将噻嗪类或类噻嗪类利尿剂作为高血压的一线单方治疗的新服用者
结局队列	发生血管性水肿或急性心肌梗死
风险暴露期间	治疗后的第一天，到暴露停止
模型	使用可变比例匹配的 Cox 比例风险模型

12.6.7 对照问题

为了评估我们的研究设计是否产生符合事实的估计，我们增加了一组对照问题，其中真实的效应大小是已知的。对照问题可分为阴性对照，危险比为 1，阳性对照，已知危险比大于 1。由于一些原因，我们使用了真实阴性对照，并在阴性对照的基础上合成阳性对照。第 18 章详细讨论了如何定义和使用对照问题。

12.7 使用 ATLAS 进行研究

在这里，我们演示了如何使用 ATLAS 中的评估功能来实现这项研究。在 ATLAS 左边的工具条点击 Estimation ，创建一个新的评估研究。确保给这项研究起一个容易辨认的名字。通过点击 按钮可以随时保存研究设计。

在评估设计功能中有三个部分：比较组、分析设置和评估设置。我们可以指定多个比较和多个分析设置，ATLAS 将以不同的分析任务执行所有这些组合。这里我们将讨论每个部分:

12.7.1 比较队列设置

一项研究可以有一个或多个比较组。点击“添加比较组”，将打开一个新的对话框。单击 以选择目标和对照队列。通过点击“添加结局”，我们可以添加两个结局队列。我们假定队列已经在 ATLAS 里被创建了，如第 10 章表述的那样。附录提供了目标队列（附录 B.2）、对照队列（附录 B.5）和结局队列（附录 B.4 和附录 B.3）的完整定义。完成后，对话框应该如图 12.6 所示。

注意，我们可以为一个目标-对照对选择多个结局。每个结局都将被独立对待，并将进行一个独立的分析。

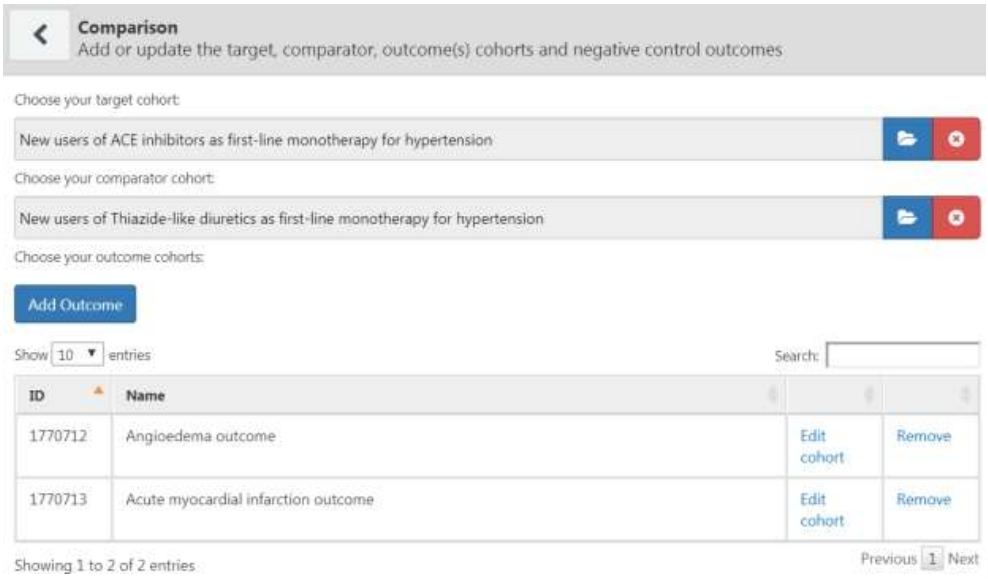


图 12.6:比较对话框

阴性对照结局

阴性对照结局是指不被认为是由目标或对照造成的结局，因此真正的危险比为 1。理想情况下，我们应该对每个结局队列有适当的队列定义。然而，通常我们只有一个概念集，每个阴性对照结局对应一个概念，以及一些标准逻辑来将这些转化为结局队列。这里我们假设概念集已经像第 18 章中描述的那样创建好了，并且可以简单地选择。阴性对照概念集应该包含每个阴性对照对应的一个概念，不包括子代。图 12.7 为本研究使用的阴性对照概念集。

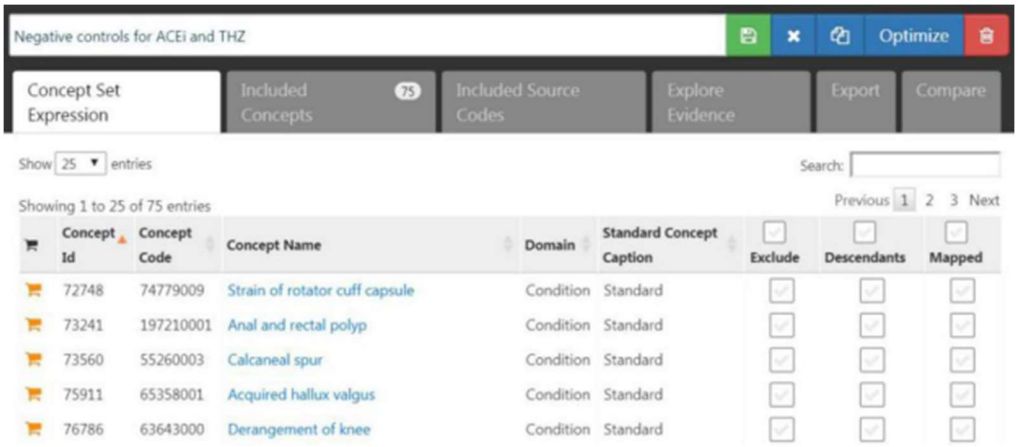


图 12.7：阴性对照概念集

概念包括

在选择要包括的概念时，我们可以指定要生成哪些协变量，例如要在倾向性模型中使用的那些。在这里指定协变量时，所有其他的协变量（除了您指定的那些）都不会被考虑。我们通常希望包括所有的基线协变量，让正则化回归建立一个模型平衡所有协变量。我们想要指定特定的协变量的唯一原因是想复现一个现有研究，而这个现有研究的协变量是手动选择的。这些包括选择可以在这个比较部分或分析部分中进行设定，因为有时它们适用于某个比较组（例如，一个比较组中已知的混杂因素），有时它们适用于分析部分（例如，在评估特定的协变量选择策略时）。

概念排除

除了指定要包括哪些概念，我们也可以指定要排除哪些概念。当我们在这个地方提交一个概念集时，我们将使用那些我们提交的之外的每一个协变量。当使用默认的协变量集时，包括了开始治疗当天所有药物和操作，我们必须排除目标和对照治疗，以及与它们直接相关的所有概念。例如，如果目标暴露是一种注射的药物，我们不仅要排除该药物，还要排除倾向性模型中的注射操作。在这个例子中，我们要排除的协变量是 ACEi 和 THZ。图 12.8 显示我们选择了一个概念集，它包含所有这些概念，包括它们的子代。

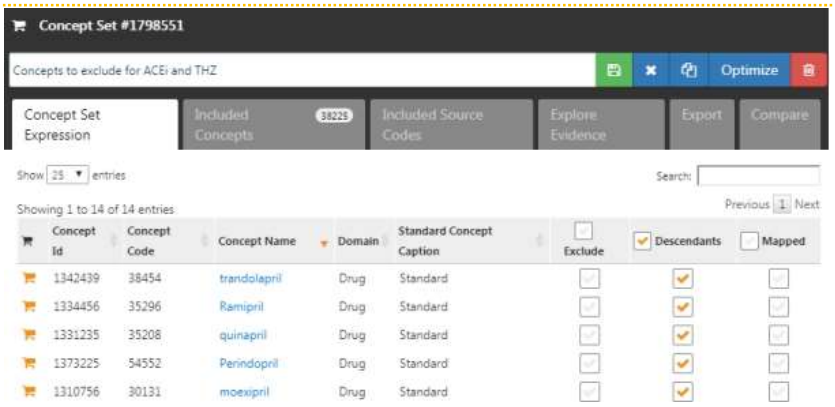


图 12.8:定义要排除概念的概念集。

选择要排除的阴性对照和协变量后，比较对话框的下半部分应该如图 12.9 所示。

图 12.9：比较窗口，显示阴性对照和排除的概念集。

12.7.2 效应估计分析设置

在关闭比较对话框后，我们可以点击“添加分析设置”。在标有“分析名称”的框中，我们可以为分析提供一个唯一的名称，以便在将来容易记住和定位。例如，我们可以将名称设置为“倾向性评分匹配”。

研究人群

有许多选项可以用来指定研究人群，即是要纳入分析的一组受试者。许多选项与队列定义工具中设计目标和对照队列的选项是重叠的。在评估设置中而不是队列定义中使用这些选项的一个原因是可重复利用性；我们可以完全独立地定义目标、对照和结局队列，并在以后添加他们之间的依赖关系。例如，如果我们希望移除在治疗开始前就发生结局的人，我们可以在目标和对照队列的定义中这样做，但我们需要随后为每个结局创建单独的队列！相反，我们可以选择在分析设置中移除具有先前结局的人，这样我们可以为我们两个目标结局（以及我们的阴性对照结局）再次使用我们的目标和对照队列。

研究的**开始和结束日期**可以用来将分析限制在一个特定的时间段。研究结束日期也会截短风险时间窗，意味着研究结束日期之后的结局将不再被考虑。选择研究开始日期的一个原因可能是被研究的药物中有一种是新的，之前并不存在。通过对问题“**将分析限制在同时存在这两种暴露的时间段**”回答“是”，可以自动对此进行调整。调整研究开始和结束日期的另一个原因可能是医疗实践随着时间的推移而改变（例如，由于药物警戒），而我们只对药物以一种特定方式处理的时期感兴趣。

选项“**只包含每个受试者的第一次暴露吗？**”可用于限制到每位患者的首次暴露。通常这已经在队

列定义中完成了，就像本例中的情况一样。类似地，选项“**在索引日期之前，一位受试者被纳入队列所需的最低连续观察时间**”通常已经在队列定义中设置，因此可以在这里保留为 0。在索引日期之前的观察时间（如 OBSERVATION_PERIOD 表中定义的那样）可以确保有足够的关于患者的信息来计算倾向性评分，并且经常用来确保患者是真正的新患者，之前没有过暴露。

“**移除同时存在于目标队列和对照队列中的受试者?**”和选项“**如果一位受试者在多个队列中，为防止重叠，当新的风险时间段开始时是否对原风险时间段进行删除?**”定义了当一位受试者同时出现在目标和对照队列中时，会发生什么。第一种设置有三个选择:

- “**保留所有**”表示在两个队列中都保留受试者。这个选项可能会重复计算受试者和结局。
- “**保留第一**”表示将受试者保留在第一个出现的队列中。
- “**移除所有**”表示从两个队列中均移除受试者。

如果选择了“保留所有”或“保留第一”选项，我们可能希望对同时处于两个队列中的受试者进行监测。如图 12.10 所示。默认情况下，风险时间段是基于队列的开始和结束日期来定义的。在本例中，风险时间段在队列开始的后一天开始，在队列结束时停止。如果不监测这两个队列的风险时间段，它们可能会重叠。如果我们选择保留所有会尤其成问题，因为任何在此重叠期间发生的结局（如图所示）将被计算两次。如果我们选择监测，第一个队列的风险时间段将在第二个队列的风险时间段开始时结束。

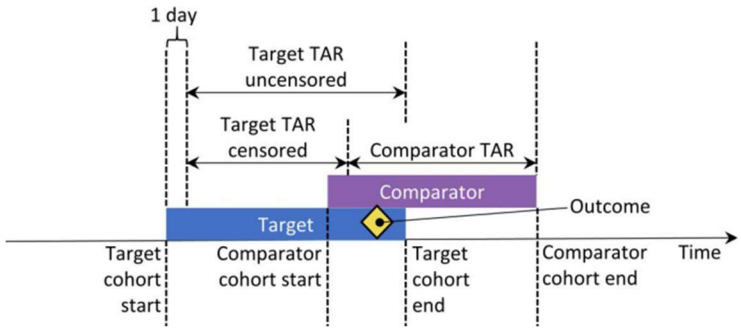


图 12.10:两组受试者的风险时间段（time-at-risk，TAR），假设风险时间段在治疗开始的后一天开始，并在暴露结束时停止。

我们可以选择**移除在风险时间窗开始之前就发生结局的受试者**，因为通常第二个结局的出现是第一个结局的延续。例如，当一个人出现心力衰竭时，很可能会出现第二次，这意味着心力衰竭可能在这两次之前并不会完全恢复。另一方面，有些结局是周期性的，患者可能会有不止一次的独立发作，比如上呼吸道感染。如果我们要移除那些之前已经出现结局的受试者，我们可以选择**在确定之前结局发生情况时应该回顾多少天。**

我们对示例研究的选择如图 12.11 所示。由于我们的目标人群和对照人群的定义已经限制在首次暴露并要求了开始治疗前需要的观察时间，因此我们不适用这些情况。

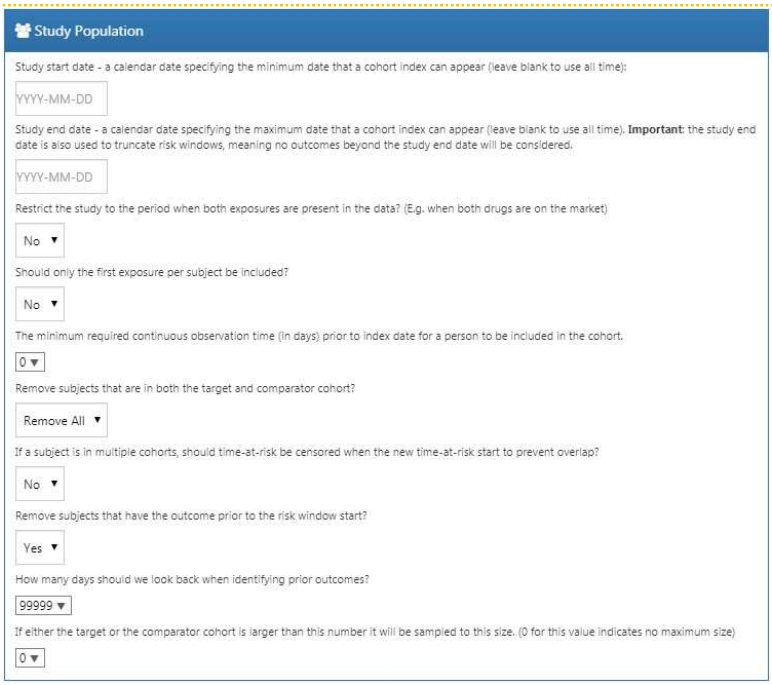


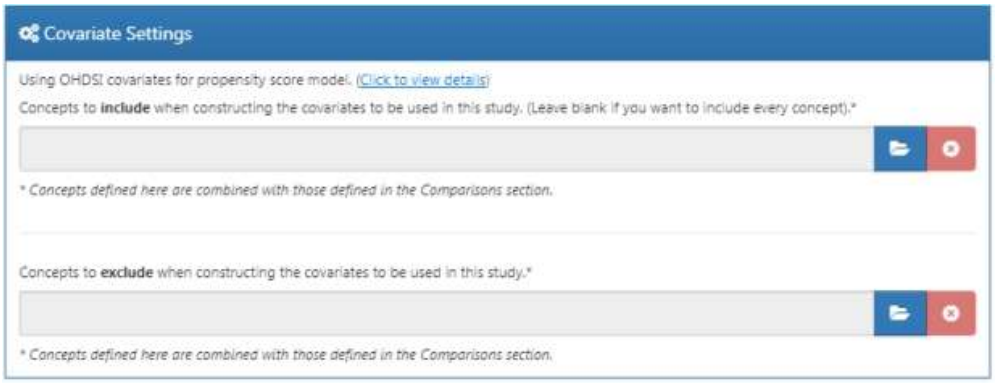
图 12.11：研究人群设置。

协变量设置

这里我们指定要构建的协变量。这些协变量通常用于倾向性评分模型，但也可以包括在预测结局的模型（Cox 比例风险模型）。如果我们**点击查看我们的协变量设置的细节**，我们可以选择构建哪些协变量集。然而，建议使用默认集，已经为人口统计、所有诊断、药物、操作、检查等构建协变量。

我们可以通过指定要包含和/或排除的概念来修改协变量集。这些设置与第 12.7.1 节中关于比较组设置的设置相同。之所以能在两个地方找到它们，是因为有时这些设置与特定的比较组有关，就像这里的情况一样，因为我们希望排除正在比较的药物。而有时这些设置与特定的分析有关，执行分析时使用特定的分析设置进行特定的比较，OHDSI 工具将取这些集合的并集。

图 12.12 显示了我们对这项研究的选择。注意，我们选择将子代添加到要排除的概念中，我们在图 12.9 中的比较设置中定义了这个概念。



图

12.12:协变量设置。

风险时间段

风险时间段是相对于我们的目标和对照队列的开始和结束日期来定义的。在我们的例子中，我们将队列起始日期设置为开始治疗时，并将队列结束日期设置为暴露停止时（至少 30 天）。我们将风险时间段的开始时间设定在队列开始后的一天，也就是开始治疗后的一天。设定风险时间开始时间晚于队列开始时间的一个原因是，如果我们不相信从生物学上讲它们可能是由药物引起，那么我们可能希望排除在开始治疗当天发生的结局事件。

我们将风险时间段的末端设定为队列末端，即暴露停止时。我们可以选择设置结束日期更后一些，比如当我们认为紧随治疗结束之后的事件仍然可以归因于暴露。在极端情况下，我们可以将风险时间段的结束时间设置为队列结束日期之后的大量天数（例如 99999），这意味着我们将有效地随访受试者直到观察结束。这样的设计有时被称为意向性（intent-to-treat）设计。

风险时间段为零的患者不会增加任何信息，因此风险时间段最小的天数通常设定为一天。如果存在一个副作用潜伏期，那么最小天数可能会设置增加以获得更有信息量的效应比例。它也可以用来创建一个类似于随机试验的队列（例如，随机试验中的所有患者都观察了至少 N 天）。

图 12.13：风险时间段设置。

一个黄金法则是在设计队列研究时永远不要使用在队列开始之后的信息来定义研究人群，因为这样的会引入偏倚。但是，如果我们要求每位受试者都至少有一年的风险时间段，我们将可能把我们的分析限制于良好耐受治疗的人群。这样的设定应当只适用于极端事例。

倾向性评分调整

在选择研究人群时，我们可以删除具有极端 PS 值的受试者。我们可以选择移除顶部和底部的百分比，或者我们可以移除那些偏好评分落在我们指定范围之外的受试者。通常不建议**修剪**队列，因为它需要丢弃观察信息，这会降低统计学效能。在某些情况下，可能需要进行修剪，例如在使用 IPTW 时。

此外，我们可以选择通过倾向性评分**分层或匹配**来代替修剪。在分层时，我们需要指定**层的数量**，以及是否根据目标、对照或整个研究人群来选择层。在匹配时，我们需要指定**对照组中与目标组每个人匹配的最大数值**。典型的值是 1，表示一比一匹配；或者是一个很大的数（例如 100）表示可变比例匹配。我们还需要指定卡尺：匹配的倾向性评分的最大容许差异。卡尺可以在不同的卡尺刻度上定义：

- **倾向性评分量表：**PS 本身

- **标准化量表：**以 PS 分布的标准差表示
- **标准化 logit 量表：**对 PS 分布进行 logit 变换后的标准差，使 PS 分布更趋于正态分布。

如果有疑问，我们建议使用默认值，或者参考 Austin(2011)关于这个主题的工作。

拟合大规模倾向性模型的计算成本可能很高，因此我们可能希望将用于拟合模型的数据限制为数据的一个样本。默认目标和对照队列的最大规模设置为 250,000。大多数研究不会达到这一限制，更多的数据也不太可能产生更好的模型。注意，虽然可以使用数据样本来拟合模型，但模型将用于计算整个总体的 PS。

测试每个协变量与目标分配的相关性? 如果有任何协变量具有异常高的相关性（无论是正还是负），将会导致错误。这一步避免了冗长的倾向性模型计算却只得到了极端的概率分离的情况。找到非常高的单变量相关性，可以让您回顾协变量，以确定为什么它具有高相关性，以及是否应该删除它。

在拟合模型时使用正则化?标准的过程是在倾向性模型中包含许多协变量(通常超过 10,000 个)。为了适应这些模型，需要一些正则化。如果只包含几个精心挑选的协变量，也可以不经过正则化就能拟合模型。

图 12.14 显示了我们对这项研究的选择。注意，我们通过将最大匹配人数设置为 100 来选择可变比例匹配。

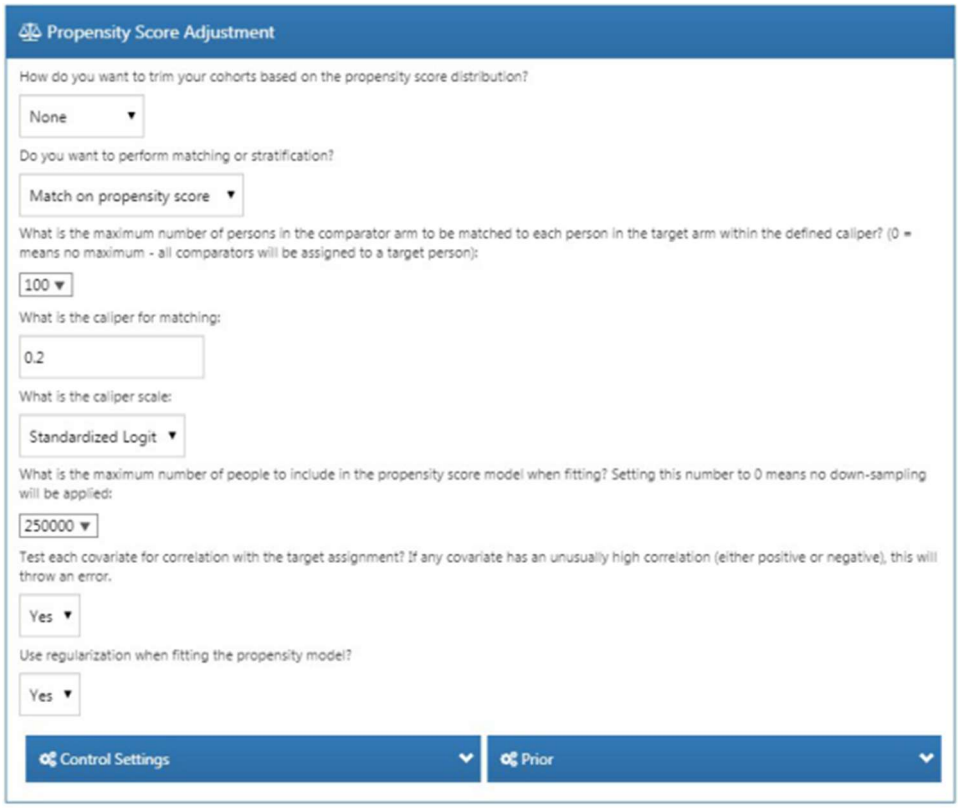


图 12.14:倾向评分调整设置。

结局模型设置

首先，我们需要**指定统计模型，我们将使用它来估计目标和对照队列之间结局的相对风险。**我们可以在 Cox、Poisson 和 Logistic 回归之间进行选择，如第 12.1 节所简要讨论的那样。在我们的例子中，我们选择了 Cox 比例风险模型，该模型考虑到第一个事件发生的时间与可能的监测。接下来，我们需要明确回归**是否应该以层为条件。**理解条件作用的一种方法是，设想在每一层产生一个单独的估计，然后在各层之间进行合并。对于一比一匹配，这可能是不必要的，而且会降低效能。对于分层或可变比例匹配，这是必需的。

我们也可以选择**将协变量添加到结局模型中**来校正分析。这可以通过添加或使用倾向性模型来实现。然而，尽管通常有足够的数据来拟合倾向性模型，两个治疗组虽然都有许多受试者，但通常仅有很少的数据来拟合结局模型，因为只有少数受试者发生了结局。因此，我们建议保持结局模型尽可能简单，不包括额外的协变量。

我们也可以**选择使用逆处理概率加权（IPTW）**来代替对倾向性评分进行分层或匹配。

如果我们选择在结局模型中包含所有的协变量，那么当有许多协变量时，在拟合模型时使用正则化是有意义的。注意，为实现无偏估计，不能将正则化应用于处理变量。

图 12.15 显示了我们对这项研究的选择。因为我们使用可变比例匹配，所以我们必须对层（即匹配集）的回归设定条件。

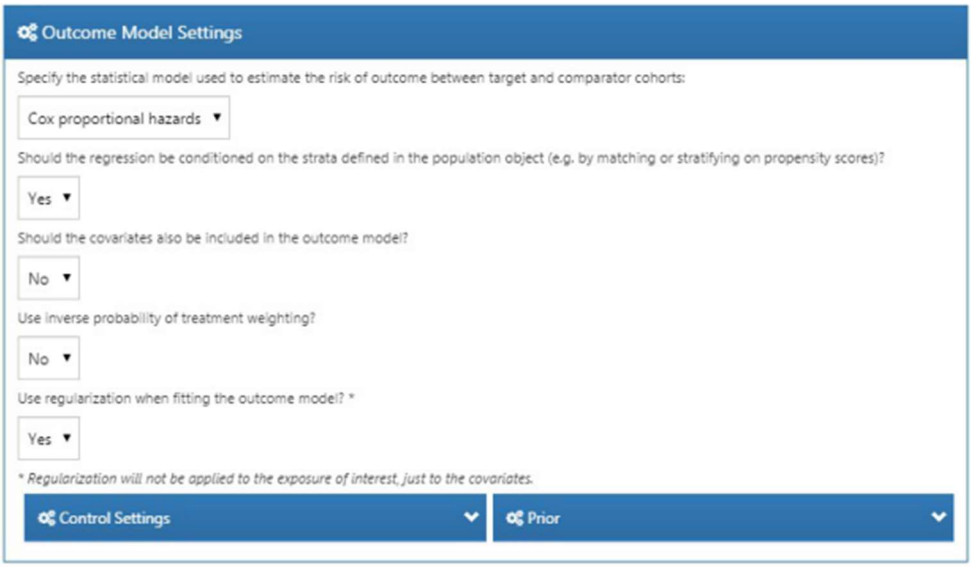


图 12.15：结局模型设置。

12.7.3 评估设置

如 18 章所述，阴性和阳性对照都应纳入我们的研究中，以评估受试者特征并进行经验性校准。

阴性对照结局队列定义

在 12.7.1 节中 我们选择了代表阴性对照结果的概念集。但是，我们需要逻辑将概念转换为队列，以用作分析结果。ATLAS 提供三种可供选择的标准逻辑。第一个选择是使用**全部出现**还是仅使用**首次**

出现的概念。第二个选择确定**是否应考虑随后概念的出现。**例如，后代“脚上的内生指甲”的出现也可以算作祖先的“内生指甲”。第三种选择指定在寻找概念时应考虑哪些领域。

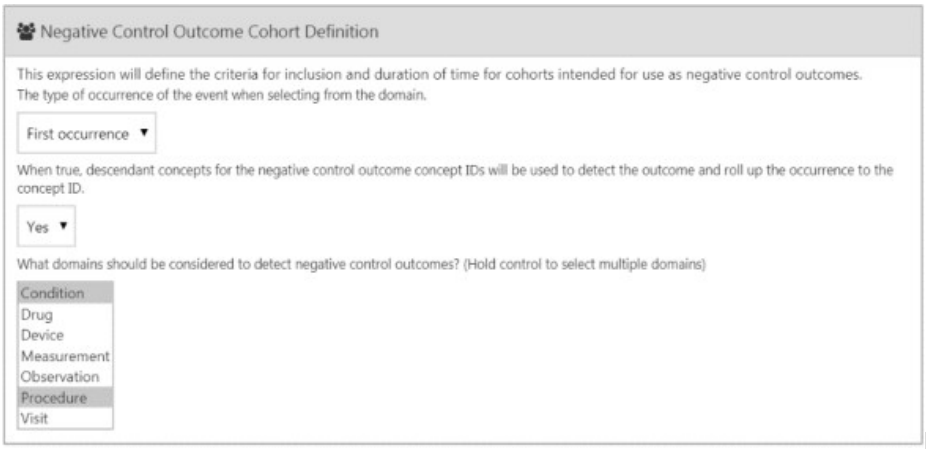


图 12.16：阴性对照结果队列定义设置。

阳性对照的合成

除阴性对照外，我们还可以包括阳性对照，它们是暴露结果对，其中存在因果效应以及已知效应大小。由于种种原因，真正的阳性对照存在问题，因此我们依赖于合成的阳性对照，这些阴性对照衍生自18.章所述的阴性对照。我们可以选择执行阳性对照合成。如果为“是”，则必须选择模型类型，为“泊松”和“生存”模型。由于我们在评估研究中使用了 Cox 模型，因此我们应该选择“生存”模型。我们将阳性对照合成的风险时间模型定义为与我们的估计设置相同，并且类似地模拟暴露前最低要求在暴露前连续观察的选择，仅考虑第一次暴露，也仅纳入第一个结局，并剔除已有结局的受试者。图 12.15 显示了阳性对照合成的设置。

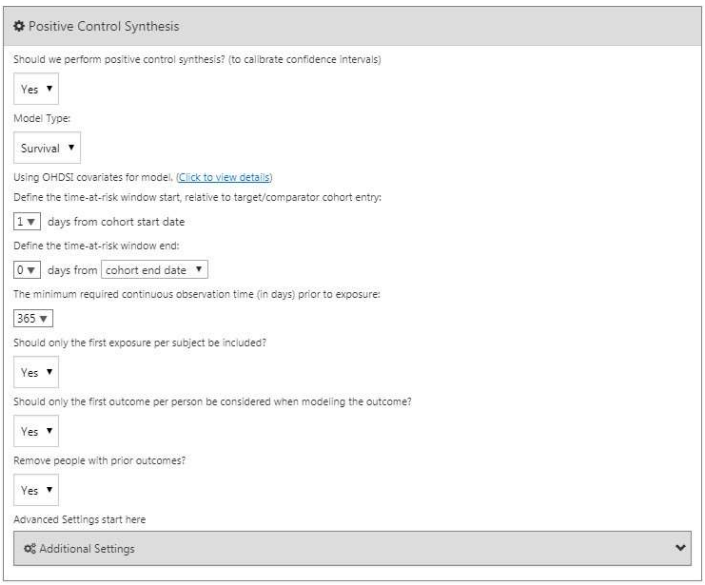


图 12.17：阴性对照结局队列定义设置。

12.7.4 运行研究套件

既然我们已经完全定义了我们的研究，我们可以将其导出为可执行的 R 软件包。这个研究包包含了在执行该研究所需的一切。这包括可用于实例化目标组、对照组和结局队列的队列定义、阴性对照概念集和创建阴性对照结果队列的逻辑，以及执行分析的 R 代码。在生成软件包之前，请确保保存您的研究，然后单击 Utility 键。在这里，我们可以回顾将要执行的一组分析。如前所述，比较和分析设置的每一种组合都会导致一次分析。在我们的例子中，我们指定了两次分析:治疗急性心肌梗死（AMI）的 ACEi 和 THZ，以及 ACEi 和 THZ 产生的血管水肿，两者都使用倾向性评分匹配。

我们必须为我们的软件包提供一个名称，之后我们可以点击“下载”来下载压缩文件。zip 文件包含一个 R 包，它具有 R 包通常需要的文件夹结构。(Wickham, 2015)要使用这个包，我们建议使用 R Studio。如果您正在本地运行 R Studio，请解压缩文件，然后双击.Rproj 文件打开。如果你在 R studio 服务器上运行 R Studio，请单击 上载并解压缩文件，然后点击.Rproj 文件打开项目。

在 R Studio 中打开项目后，可以打开 README 文件，并按照说明进行操作。请确保将所有文件路径更改为系统上的现有路径。

运行研究时可能出现的一个常见错误信息是“协变量和检测到的治疗之间的高度相关性”。这表明当拟合倾向性模型时，观察到一些协变量与暴露高度相关。请检查错误信息中提到的协变量，如果发现，将它们从协变量集中(请参阅 12.1.2 节)。

12.8 使用 R 执行研究

我们也可以自己编写代码，而不是用 ATLAS 来编写执行研究的代码。我们想这么做的一个原因可能是，相对于 ATLAS，R 提供了更大的灵活性。例如，如果我们希望使用定制协变量或线性结局模型，我们需要编写一些定制的代码，并将其与 OHDSI R 包提供的功能相结合。

对于我们的示例研究,我们将依赖 CohortMethod 包来执行我们的研究。CohortMethod 从 CDM 数据库中提取必要的数据，并可以使用多组协变量来构建倾向性模型。在下面的例子中，我们首先只考虑血管性水肿作为结局。在第 12.8.6 节中，我们将描述如何将其扩展至包括急性心肌梗塞（AMI）和阴性对照结果。

12.8.1 队列实例化

我们首先需要实例化目标和结局队列。第 10 章描述了如何实例化队列。附录提供了目标(附录 B.2)、比较(附录 B.5)和结局(附录 B.4)队列的完整定义。我们将假设 ACEi、THZ 和血管性水肿队列已在一个名为 scratch.my_cohorts 的表格中实例化，队列定义分别为 IDs 1、2 和 3。

12.8.2 数据提取

我们首先需要告诉 R 如何连接到服务器。CohortMethod 使用 DatabaseConnector 包。

它提供了一个名为 createConnectionDetails 的功能。Type?createConnectionDetails 为各种数据管理系统（DBMS）所需的特定设置。比如可以使用以下代码连接到 PostgreSQL 数据库：

```
library(CohortMethod)
connDetails <- createConnectionDetails(dbms = "postgresql",
                                       server = "localhost/ohdsi",
                                       user = "joe",
                                       password = "supersecret")

cdmDbSchema <-"my_cdm_data" cohortDbSchema <- "scratch"
cohortTable <- "my_cohorts" cdmVersion <- "5"
```

最后四行定义了 cdmDbSchema、cohortDbSchema 和 cohortTable 变量以及 CDM 版本。稍后我们将使用这些信息告诉 R CDM 格式的数据在哪里存在，相关队列在哪里创建，以及使用什么版本的 CDM。请注意，对于 Microsoft SQL Server，数据库架构需要同时指定数据库和架构。

例如 cdmDbSchema <- " my_cdm_data.dbo"。

现在我们可以告诉 CohortMethod 提取队列，构造协变量并提取所有必要的数据以供我们分析：

```
# target and comparator ingredient concepts:
aceI <- c(1335471,1340128,1341927,1363749,1308216,1310756,1373225, 1331235,1334456,1342439)
thz <- c(1395058,974166,978555,907013)


# Define which types of covariates must be constructed:
cs<-createDefaultCovariateSettings(excludedCovariateConceptIds=c(aceI,thz),
                                   addDescendantsToExclude =TRUE)

#Load data:
cmData<-getDbCohortMethodData(connectionDetails=connectionDetails,
                              cdmDatabaseSchema=cdmDatabaseSchema,
                              oracleTempSchema = NULL,
                              targetId = 1,
                              comparatorId = 2,
                              outcomeIds = 3,
                              studyStartDate="",
                              studyEndDate ="",
                              exposureDatabaseSchema=cohortDbSchema,
                              exposureTable = cohortTable,
                              outcomeDatabaseSchema = cohortDbSchema,
                              outcomeTable =cohortTable,
                              cdmVersion = cdmVersion,
                              firstExposureOnly = FALSE,
                              removeDuplicateSubjects = FALSE,
                              restrictToCommonPeriod = FALSE,
                              washoutPeriod = 0,
                              covariateSettings =cs)
cmData
```

```
## CohortMethodData object
##
## Treatment concept ID: 1
 ## Comparator concept ID: 2
## Outcome concept ID(s): 3
```

有很多参数，但是它们都记录在 CohortMethod 手册里。createDefaultCovariateSettings 函数在特征提取包进行了描述。简而言之，我们将函数指向包含队列的工作表，并指定该表中哪些队列定义 ID 标识了目标、比较和结局。我们指示应构建默认的协变量集，包括所有条件，药物暴露以及在索引日期或之前找到的程序的协变量。如本节所述 12.1，我们必须从协变量集中排除目标和比较治疗，在这里我们通过列出两个类别中的所有成分来实现，并告诉 FeatureExtraction 也排除所有子目录，从而排除包含这些成分的所有药物。

所有关于队列、结果和协变量的数据均从服务器中提取并存储在 cohortMethodData 的对象中。该对象使用软件包 ff 来存储信息，以确保即使数据很大，R 也不会耗尽内存，如第 8.4.2 节所述。

我们可以使用 generic summary（）函数来查看提取的数据的更多信息：

```
summary(cmData)
```

```
## CohortMethodData object summary ##
## Treatment concept ID: 1 ## Comparator concept ID: 2 ## Outcome concept
ID(s): 3 ##
## Treated persons: 67166 ## Comparator persons: 35333 ##
## Outcome counts:
##  Event count Person count ## 3      980 891
##
## Covariates:
## Number of covariates: 58349
## Number of non-zero covariate values: 24484665
```

创建 cohortMethodData 文件可能需要花费大量的计算时间，将其保存以备将来使用可能是一个好主意。因为 cohortMethodData 使用软件包 ff，所以我们不能使用 R 的常规保存功能。相反，我们必须使用 saveCohortMethodData（）函数：

```
saveCohortMethodData(cmData, "AceiVsThzForAngioedema")
```

我们可以使用 loadCohortMethodData（）函数在之后的程序中加载数据。

定义新用户

通常，将新用户定义为首次使用药物（目标或比较药物），通常使用洗脱期（首次使用前的最少天数）来增加其真正首次使用的可能性。使用 CohortMethod 程序包时，可以通过以下三种方式强制实施首次起始用药的必要要求：

1. 在定义队列时。
2. 使用 getDbCohortMethodData 函数加载队列时，可以使用 firstExposureOnly，removeDuplicateSubjects，restrictToCommonPeriod 和 washoutPeriod 参数。
3. 使用 createStudyPopulation 函数定义研究人群时（请参见下文）。

选项 1 的优点是，在 CohortMethod 程序包之外已经完全定义了输入队列，并且可以在同一队列上使用外部队列表征工具。 选项 2 和 3 的优点是，它们使您免去了首次使用限制的麻烦，例如，允许您直接使用 CDM 中的 DRUG_ERA 表。选项 2 比 3 更有效，因为仅会提取首次使用的数据，而选项 3 效率较低，但允许您将原始队列与研究人群进行比较。

12.8.3 定义研究人群

通常，暴露队列和结果队列将彼此独立定义。当我们希望产生一个效应量估计值时，我们需要进一步限制这些队列并将它们放在一起，例如，通过移除暴露前具有结局的暴露对象，并仅将结果保持在定义的风险范围内。为此，我们可以使用 createStudyPopulation 函数：

```
studyPop <- createStudyPopulation(cohortMethodData = cmData,
                                  outcomeId = 3,
                                  firstExposureOnly = FALSE,
                                  restrictToCommonPeriod = FALSE,
                                  washoutPeriod = 0,
                                  removeDuplicateSubjects = "remove all",
                                  removeSubjectsWithPriorOutcome =TRUE,
                                  minDaysAtRisk = 1,
                                  riskWindowStart = 1,
                                  startAnchor = "cohort start",
                                  riskWindowEnd = 0,
                                  endAnchor = "cohort end")
```

注意我们已将 firstExposureOnly 和 removeDuplicateSubjects 设置为 FALSE，将 washoutPeriod 设置为 0，因为我们已经在队列定义中应用了这些条件。我们指定将要使用的结果 ID，并且删除风险窗口开始日期之前已出现结局的受试者。风险窗口定义为在队列开始日期之后的第二天开始（riskWindowStart = 1 和 startAnchor = “队列开始”），并且该风险窗口在队列暴露结束时结束（riskWindowEnd = 0 和 endAnchor = “队列结束”），在队列中定义为暴露结束。请注意，风险窗口会在观察结束或研究结束日期时自动被截断。我们还将删除没有时间风险的受试者。要查看研究人群中还剩下多少人，我们可以使用 getAttritionTable 函数：

```
getAttritionTable(studyPop)
```

#		description	targetPersons	comparatorPersons	.
#	1	Original cohorts	67212	35379	.
#	2	No prior outcome	67166	35333	.
#	3		67061	35238	.
#	4	Have at least 1 days at	66780	35086	.

risk

12.8.4 倾向性评分

我们可以使用由 getDbcohortMethodData（）构造的协变量来拟合倾向性模型，并为每个人计算 PS：

```
ps <- createPs(cohortMethodData = cmData, population = studyPop)
```

createPs 函数使用 Cyclops 包以适合大规模正则逻辑回归。为了拟合倾向性模型，Cyclops 需要知道指定先验方差的超参数值。默认情况下，Cyclops 将使用交叉验证来估计最佳超参数。但是，请注意，这可能需要很长时间。您可以使用 createPs 函数的 Priority 和 Control 参数来指定 Cyclops 的行为，包括使用多个 CPU 来加快交叉验证的速度.

在这里，我们使用 PS 执行可变比率匹配：

```
matchedPop <- matchOnPs(population = ps, caliper = 0.2,
                        caliperScale = "standardized logit", maxRatio = 100)
```

或者，我们可以调用 trimByPs，trimmByPsToEquipoise 或 stratifyByPs 函数。

12.8.5 结局模型

结局模型是描述哪些变量与结局关联的模型。在严格的假设下，治疗变量的系数可以解释为因果关系。在这种情况下，我们拟合了 Cox 比例风险模型，该模型以匹配组为条件（分层）：

```
outcomeModel <- fitOutcomeModel(population = matchedPop,
                                modelType = "cox",
                                stratified = TRUE)
outcomeModel
```

```
## Model type: cox
## Stratified: TRUE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE ## Status: OK
##
##                Estimate lower .95 upper .95   logRr   seLogRr
## treatment        4.3203    2.4531     8.0771    1.4633   0.304
```

12.8.6 运行多个分析

通常，我们想要执行多个分析，例如对包括阴性对照在内的多种结果进行分析。CohortMethod

提供有效执行此类研究的功能。这将在运行多个分析的程序包小插图. 简要地说，假设已经建立了感兴趣的结果和阴性对照队列，我们可以指定我们要分析的所有目标-比较-结局组合：

```
aceI <- c(1335471,1340128,1341927,1363749,1308216,1310756,1373225,
          1331235,1334456,1342439)
thz <- c(1395058,974166,978555,907013)

cs <- createDefaultCovariateSettings(excludedCovariateConceptIds = c(aceI,thz),
                                     addDescendantsToExclude = TRUE)

cmdArgs <- createGetDbCohortMethodDataArgs( studyStartDate = "",
  studyEndDate = "", firstExposureOnly =FALSE,
  removeDuplicateSubjects = FALSE,
  restrictToCommonPeriod = FALSE,
  washoutPeriod = 0,
  covariateSettings =cs)
spArgs <-
  createCreateStudyPopulationArgs( firstExp
  osureOnly = FALSE,
  restrictToCommonPeriod = FALSE,
  washoutPeriod = 0,
  removeDuplicateSubjects = "remove all",
  removeSubjectsWithPriorOutcome = TRUE,
tcosList <- list(tcos)
```

接下来，我们指定在调用先前示例中描述的各种函数时应选用哪些参数，并得出结果：

```
   minDaysAtRisk = 1,
   startAnchor="cohortstart",
   addExposureDaysToStart = FALSE,
   endAnchor = "cohort end",
   addExposureDaysToEnd=TRUE)

psArgs <- createCreatePsArgs()

matchArgs <- createMatchOnPsArgs(
 caliper = 0.2,
 caliperScale = "standardized logit", maxRatio = 100)
fomArgs <- createFitOutcomeModelArgs( modelType = "cox",
  stratified =TRUE)
```

然后，我们将它们组合到单一分析设置对象中，该对象将提供唯一的分析 ID 和说明。我们可以将一个或多个分析设置对象合并到一个列表中：

```
cmAnalysis <- createCmAnalysis( analysisId = 1,
  description = "Propensity score matching",
  getDbCohortMethodDataArgs = cmdArgs,
  createStudyPopArgs = spArgs,
  createPs = TRUE,
  createPsArgs = psArgs,
  matchOnPs = TRUE,
  matchOnPsArgs = matchArgs fitOutcomeModel = TRUE,
  fitOutcomeModelArgs = fomArgs)

cmAnalysisList <- list(cmAnalysis)
```

现在，我们可以运行该研究，包括所有比较和分析设：

```
result <- runCmAnalyses(connectionDetails = connectionDetails,
                        cdmDatabaseSchema = cdmDatabaseSchema,
                        exposureDatabaseSchema = cohortDbSchema,
                        exposureTable = cohortTable,
                        outcomeDatabaseSchema = cohortDbSchema,
                        outcomeTable = cohortTable,
                        cdmVersion = cdmVersion,
                        outputFolder = outputFolder,
                        cmAnalysisList =cmAnalysisList,
                        targetComparatorOutcomesList =tcosList)
```

Result 对象包含对所有已创建语句的引用。例如，我们可以检索 AMI 的结局模型：

```
omFile <- result$outcomeModelFile[result$targetId == 1 &
                                    result$comparatorId==2&
                                    result$outcomeId == 4 &
                                    result$analysisId == 1]
outcomeModel <- readRDS(file.path(outputFolder, omFile))
outcomeModel
```

```
## Model type: cox
## Stratified: TRUE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE ## Status: OK
##
##           Estimate  lower .95  upper .95    logRr  seLogRr
## treatment     1.1338    0.5921  2.1765     0.1256  0.332
```

我们还可以使用一个命令来检索所有结局的效应大小：

```
summ <- summarizeAnalyses(result, outputFolder = outputFolder)
head(summ)
```

##	analysisId	targetId	comparatorId	outcomeId	rr ...
## 1	1	1	2	72748	0.9734698 ...
## 2	1	1	2	73241	0.7067981 ...
## 3	1	1	2	73560	1.0623951 ...
## 4	1	1	2	75911	0.9952184 ...
## 5	1	1	2	76786	1.0861746 ...
## 6	1	1	2	77965	1.1439772 ...

12.9 研究成果

我们的计算仅在满足多个假设的情况下才有效。我们使用各种诊断程序来评估是否是这种情况。这些可以通过 ATLAS 产生的 R 包生成的结果中找到，也可以使用特定的 R 函数即时生成。

12.9.1 倾向性评分和模型

我们首先需要评估目标人群和比较人群是否在一定程度上具有可比性。为此，我们可以为倾向性模型计算“受试者特征性曲线下面积”（AUC）统计信息。AUC 为 1 表示根据基线协变量可以完全预测治疗方案，因此两组不可比。我们可以使用 computePsAuc 函数来计算 AUC，在我们的示例中是 0.79。使用 plotPs 函数，我们还可以生成偏好得分分布，如图 12.18 所示。在这里，我们看到，对于许多人来说，他们所接受的治疗是可以预见的，但是也有很多重叠之处，这表明可以使用调整来选择可比较的人群。

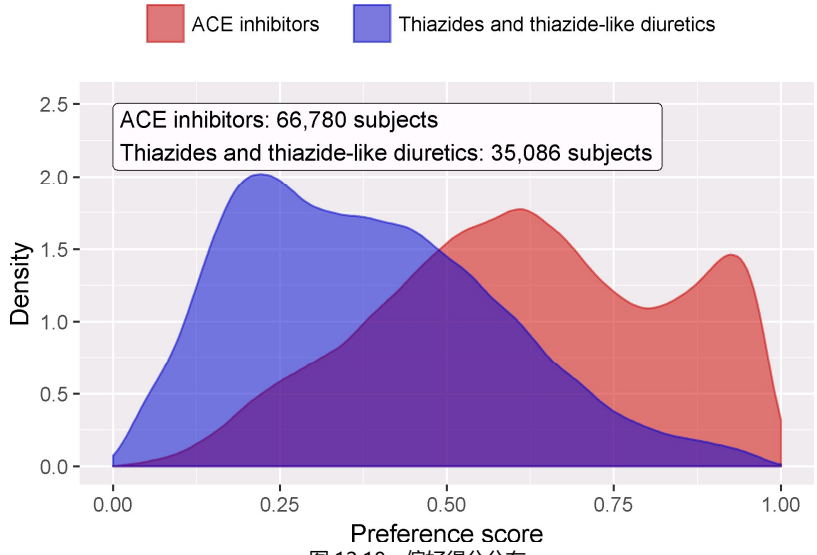


图 12.18：偏好得分分布

通常，也应检查倾向性模型本身，尤其是在模型具有高度预测性的情况下。这样，我们可能会发现哪些变量最具预测性。表 12.7 显示了我们倾向性模型中的主要预测变量。请注意，如果变量的预测性

太强，则 CohortMethod 程序包将得出信息错误，而不是尝试拟合已知的完美预测性模型。

表 12.7：ACEi 和 THZ 倾向性模型中的前 10 个预测变量。正值表示具有协变量的受试者更有可能接受目标治疗。“(截距)”表示该逻辑回归模型的截距。

Beta	Covariate
-1.42	condition_era group during day -30 through 0 days relative to index: Edema
-1.11	drug_era group during day 0 through 0 days relative to index: Potassium Chloride
0.68	age group: 05-09
0.64	measurement during day -365 through 0 days relative to index: Renin
0.63	condition_era group during day -30 through 0 days relative to index: Urticaria
0.57	condition_era group during day -30 through 0 days relative to index: Proteinuria
0.55	drug_era group during day -365 through 0 days relative to index: INSULINS AND ANALOGUES
-0.54	race = Black or African American
0.52	(Intercept)
0.50	gender = MALE

12.9.2 协变量平衡

使用 PS 的目的是使两个组具有可比性（或至少选择可比较的组）。我们必须验证是否实现了这一目标，例如，通过检查调整后基线协变量是否确实达到平衡。我们可以使用 computeCovariateBalance 和 plotCovariateBalanceScatterPlot 函数来生成图 12.19. 使用的一种经验法则是，在倾向性得分调整后，没有任何协变量的均值的绝对标准化差可以大于 0.1。在这里我们看到尽管匹配之前存在很大的不平衡，但是匹配之后我们满足了这一标准。

12.9.3 随访和效能

在拟合结局模型之前，我们可能想知道我们是否有足够的能力来检测特定的效应量。一旦确定了研究人群，就可以执行这些效能计算，因此，考虑到丢失各种纳入和排除标准（例如没有事先结果），以及由于匹配和/或修整导致的损失。我们可以使用 drawAttritionDiagram 函数查看研究中的对象损耗，如图 12.20 所示。

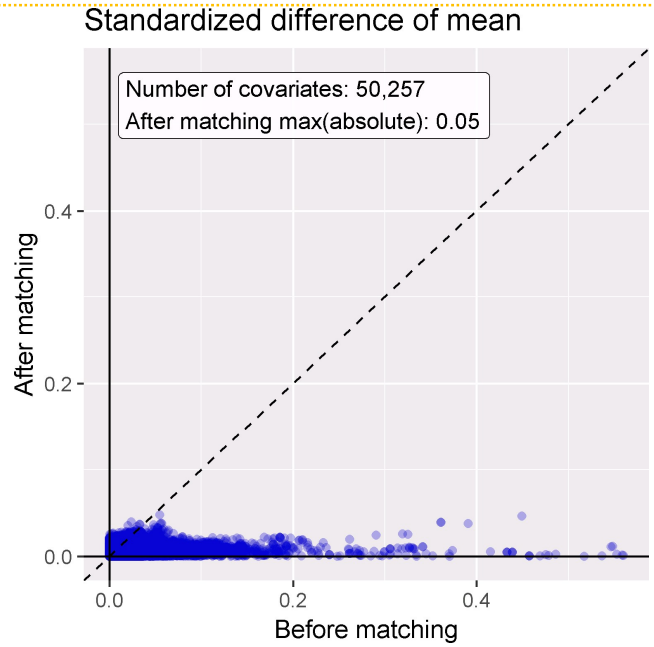


图 12.19：协变量平衡，显示了倾向性评分匹配前后均值的绝对标准化差。每个点代表一个协变量。

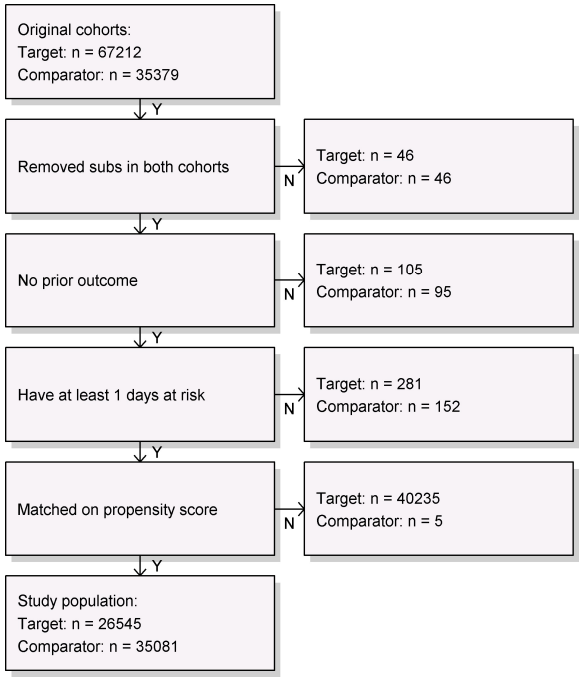


图 12.20：递减图。顶部显示的计数是满足我们的目标和比较队列定义的计数。底部的计数是进入我们的结局模型的计数，在这种情况下采用 Cox 回归。

由于在回顾性研究中样本量是固定的（数据已被收集），并且真正的效应量未知，因此，在给定预期效应量的情况下，计算效能的意义较小。而是，CohortMethod 软件包提供了 computeMdrr 函数来计算最小可检测相对风险（MDRR）。在我们的示例研究中，MDRR 为 1.69。

为了更好地了解可用的随访量，我们还可以检查随访时间的分布。我们将随访时间定义为有风险的时间，因此不根据结局的发生进行检查。getFollowUpDistribution 可以提供一个简单的概述，如图 12.21 所示，这表明两个队列的随访时间相当。

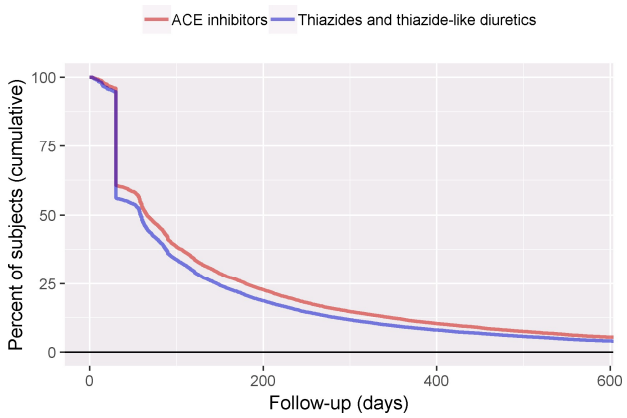


图 12.21：目标人群和比较人群的随访时间分布。

12.9.4 Kaplan-Meier 图

最后一项检查是检查 Kaplan-Meier 图，显示两个队列随时间的生存情况。使用 plotKaplanMeier 函数，我们可以创建图 12.22。

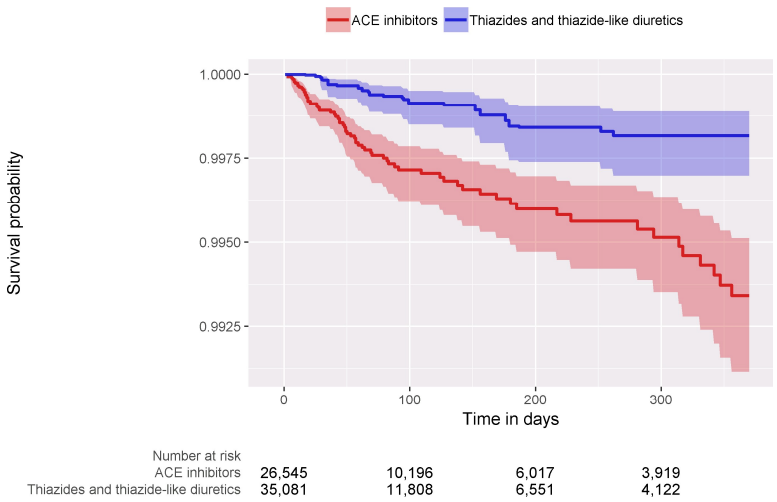


图 12.22：Kaplan-Meier 图

例如，我们可以检查我们的危险性比例假设是否成立。Kaplan-Meier 图通过 PS 自动调整分层或加权。在这种情况下，因为使用了可变比率匹配，所以调整了比较组的生存曲线，以模拟目标组在暴露给比较组的生存曲线。

12.9.5 效应大小估算

我们观察到血管性水肿的风险比为 4.32（95%置信区间：2.45-8.08），这表明与 THZ 相比，ACEi 似乎会增加血管性水肿的风险。同样，我们观察到 AMI 的危险比为 1.13（95%置信区间：0.59-2.18），表明对 AMI 几乎没有影响。如前所述，我们的诊断程序毫无疑问。但是，最终该证据的质量以及我们是否选择信任它，取决于 14 章所述的研究诊断程序未涵盖的多种因素。

12.10 总结

人群水平估计的目的是从观察数据推断因果关系。

反事实，即如果受试者接受替代暴露或非暴露，则无法观察到。

不同的设计旨在以不同的方式构造反事实。

OHDSI 方法库中实现的各种设计提供了诊断程序，以评估是否已满足创建适当反事实的假设。

12.11 练习题

准备工作

对于这些练习，我们假设已按照第 8.4.5 节中的说明安装了 R、R-Studio 和 Java。还需要 SqlRender, DatabaseConnector, Eunomia 和队列方法软件包，可以使用以下方法安装：

```
install.packages(c("SqlRender", "DatabaseConnector", "devtools"))
devtools::install_github("ohdsi/Eunomia", ref = "v1.0.0")
devtools::install_github("ohdsi/CohortMethod")
```

Eunomia 程序包在 CDM 中提供了一个模拟数据集，该数据集将在本地 R 程序中运行。可以使用以下方法获取连接详细信息：

```
connectionDetails <- Eunomia::getEunomiaConnectionDetails()
```

CDM 数据库架构师“主要”的。这些练习还利用了一些队列。Eunomia 包中的 createCohorts 函数将在 COHORT 表中创建这些：

```
connectionDetails <- Eunomia::getEunomiaConnectionDetails()
```

问题定义

与双氯芬酸的初次使用者相比，塞来昔布的初次使用者中胃肠道（GI）出血的风险是多少？

celecoxib 初次使用者队列为 COHORT_DEFINITION_ID = 1。双氯芬酸初次使用者队列为 COHORT_DEFINITION_ID = 2。GI 出血队列为 COHORT_DEFINITION_ID = 3。塞来昔布和双氯芬酸的 ID 分别为 1118084 和 1124300。风险时间从治疗当天开始，到观察结束时停止（所谓的治疗意向分析）。

练习 12.1.使用 CohortMethod R 包，使用默认的协变量集，并从 CDM 中提取 CohortMethodData。创建 CohortMethodData 的摘要。

练习 12.2.使用 createStudyPopulation 函数创建研究人群，人群纳入需要 180 天的洗脱期，排除之前有结局的受试者，并删除两个队列中都出现的受试者。我们是否丢失了受试者？

练习 12.3.在不进行任何调整的情况下拟合 Cox 比例风险模型。如果这样做会出什么问题？

练习 12.4.拟合倾向性模型。两组是否具有可比性？

练习 12.5.使用 5 个层次执行 PS 分层。是否实现了协变量平衡？

练习 12.6.使用 PS 分层拟合 Cox 比例风险模型。为什么结果与未经调整的模型不同？

参考答案可以在附录 E.8.中找到。

参考文献

1. Austin, Peter C. 2011. “Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies.” Pharmaceutical Statistics 10 (2): 150–61.
2. Farrington, C. P. 1995. “Relative incidence estimation from case series for vaccine safety evaluation.” Biometrics 51 (1): 228–35.
3. Farrington, C. P., Karim Anaya-Izquierdo, Heather J. Whitaker, Mounia N. Hocine, Ian Douglas, and Liam Smeeth. 2011. “Self-Controlled Case Series Analysis with Event-Dependent Observation Periods.” Journal of the American Statistical Association 106 (494): 417–26. https://doi.org/10.1198/jasa.2011.ap10108.
4. Hernan, M. A., S. Hernandez-Diaz, M. M. Werler, and A. A. Mitchell. 2002. “Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology.” Am. J. Epidemiol. 155 (2): 176–84.
5. Hernan, M. A., and J. M. Robins. 2016. “Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.” Am. J. Epidemiol. 183 (8): 758–64。
6. Maclure, M. 1991. “The case-crossover design: a method for studying transient effects on the risk of acute events.” Am. J. Epidemiol. 133 (2): 144–53。
7. Magid, D. J., S. M. Shetterly, K. L. Margolis, H. M. Tavel, P. J. O’Connor, J. V. Selby, and P. M. Ho. 2010. “Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blockers as second-line therapy for hypertension.” Circ Cardiovasc Qual Outcomes 3 (5): 453–58.
8. Powers, B. J., R. R. Coeytaux, R. J. Dolor, V. Hasselblad, U. D. Patel, W. S. Yancy, R. N. Gray, R. J. Irvine, A. S. Kendrick, and G. D. Sanders. 2012. “Updated report on comparative effectiveness of ACE inhibitors, ARBs, and direct renin inhibitors for patients with essential hypertension: much more data, little new information.” J Gen Intern Med 27 (6): 716–29.
9. Rassen, J. A., A. A. Shelat, J. Myers, R. J. Glynn, K. J. Rothman, and S. Schneeweiss. 2012. “One-to-many propensity score matching in cohort studies.” Pharmacoepidemiol Drug Saf 21 Suppl 2 (May): 69–80.
10. Rosenbaum, P., and D. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” Biometrika 70 (April): 41–55. https://doi.org/10.1093/biomet/70.1.41.
11. Rubin, Donald B. 2001. “Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation.” Health Services and Outcomes Research Methodology 2 (3-4): 169–88.
12. Ryan, P. B., M. J. Schuemie, and D. Madigan. 2013. “Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system.” Drug Saf 36 Suppl 1 (October): 95–106.
13. Sabroe, R. A., and A. K. Black. 1997. “Angiotensin-converting enzyme (ACE) inhibitors and angio-oedema.” Br. J. Dermatol. 136 (2): 153–58.

14. Schneeweiss, S. 2018. “Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects.” Clin Epidemiol 10: 771–88.
15. Simpson, S. E., D. Madigan, I. Zorych, M. J. Schuemie, P. B. Ryan, and M. A. Suchard. 2013. “Multiple self-controlled case series for large-scale longitudinal observational databases.” Biometrics 69 (4): 893–902.
16. Suchard, M. A., S. E. Simpson, Ivan Zorych, P. B. Ryan, and David Madigan. 2013. “Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model.” ACM Trans. Model. Comput. Simul. 23 (1): 10:1–10:17. https://doi.org/10.1145/2414416.2414791.
17. Suissa, S. 1995. “The case-time-control design.” Epidemiology 6 (3): 248–53.
18. Tian, Y., M. J. Schuemie, and M. A. Suchard. 2018. “Evaluating large-scale propensity score performance through real-world and synthetic data experiments.” Int J Epidemiol 47 (6): 2005–14.
19. Toh, S., M. E. Reichman, M. Houstoun, M. Ross Southworth, X. Ding, A. F. Hernandez, M. Levenson, et al. 2012. “Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system.” Arch. Intern. Med. 172 (20): 1582–9.
20. Vandenbroucke, J. P., and N. Pearce. 2012. “Case-control studies: basic concepts.” Int J Epidemiol 41 (5): 1480–9.
21. Walker, Alexander M, Amanda R Patrick, Michael S Lauer, Mark C Hornbrook, Matthew G Marin, Richard Platt, Véronique L Roger, Paul Stang, and Sebastian Schneeweiss. 2013. “A Tool for Assessing the Feasibility of Comparative Effectiveness Research.” Comp Eff Res 3: 11–20.
22. Whelton, P. K., R. M. Carey, W. S. Aronow, D. E. Casey, K. J. Collins, C. Dennison Himmelfarb, S. M. DePalma, et al. 2018. “2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines.” Circulation 138 (17): e426–e483.
23. Whitaker, H. J., C. P. Farrington, B. Spiessens, and P. Musonda. 2006. “Tutorial in biostatistics: the self-controlled case series method.” Stat Med 25 (10): 1768–97.
24. Wickham, Hadley. 2015. R Packages. 1st ed. O’Reilly Media, Inc.
25. Zaman, M. A., S. Oparil, and D. A. Calhoun. 2002. “Drugs targeting the renin-angiotensin-aldosterone system.” Nat Rev Drug Discov 1 (8): 621–36.

第十三章　患者水平预测

章节负责人：Peter Rijnbeek 与 Jenna Reps

临床决策是一项复杂的任务，临床医生必须基于患者的可用病史和当前的临床指南推断出诊断结果或治疗路径。现已开发出的临床预测模型可以支持上述决策过程，并已在临床实践中得到广泛应用。这些模型根据患者的组合特征，如人口统计信息、疾病史和治疗史等，来预测诊断或预后结果。

在过去的十年里，描述临床预测模型的论文数量迅速增长。目前大多数模型使用小型数据集去做预测，并且仅考虑小部分病人特征。由于低样本量导致的低统计效力，数据分析人员需要提出较强的建模假设。有限的病人特征刻画的选择依赖于现有专家知识的指导。这与现代医学的实际情况形成了鲜明对比，在现代医学中，已经形成了丰富的患者数字资料，这些远远超过了任何医疗从业者完全吸收的能力。目前，医疗保健系统正在大量生成能够存储在电子健康档案（EHR）中的患者信息。这些信息包括诊断、药物、化验结果在内的结构化数据，和包含在临床描述中的非结构化数据。通过整合大量来自患者完整 EHR 数据能获取的预测精准度尚未可知。

随着针对大型数据集的机器学习技术的进步，人们产生了使用此类数据进行患者水平预测的兴趣。但是，许多已经发表的患者水平预测领域的研究并没有遵循模型开发指南，无法进行广泛的外部验证，或者由于提供的模型细节不够，从而限制了独立研究者重现模型并进行外部验证的能力。这就导致很难公平地评价模型预测的性能，且降低了模型在临床实践中被合理运用的可能性。为了提高标准，已有一些文献详细描述了预测模型开发和发表的最佳实践指导。例如，针对个人预后或诊断多变量预测模型的透明报告声明（TRIPOD 指南）提供了报告预测模型开发和验证的明确建议，并且阐述了一些与报告透明度有关的重要事项。

基于 OHDSI，大规模针对患者个体的预测模型已成为现实，通用数据模型（CDM）可以通过前所未有的方式进行统一透明的数据分析。不断增长的 CDM 标准化数据库网络使得在全球范围内不同医疗场所中进行外部模型验证成为可能。我们相信这很快就可以为改善患者照护质量提供机会。这些模型可以提供真正个体化的医疗服务，大幅度改善患者预后。

在本章中，我们将介绍 OHDSI 用于患者水平预测的标准化框架，（Resp et al., 2018）并讨论如何使用 R 软件包 PatientLevelPrediction 高效实施现有模型的开发和验证。首先我们概要介绍在患者水平预测模型开发和评价背后的必要理论和机器学习模型实践；然后我们讨论一个具体预测问题的例子，并介绍概念，逐步指导 ATLAS 或 R 自编程的实现；最后，我们讨论如何运用 Shiny 应用程序进行研究结果的传输。

13.1 预测问题

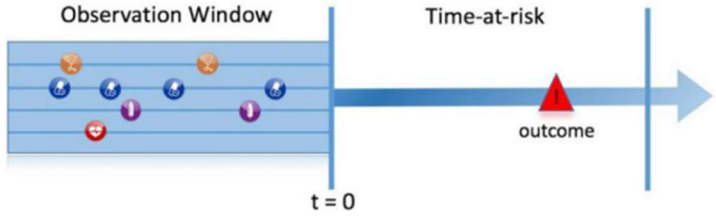


图 13.1 预测问题

图 13.1 说明了我们要解决的预测问题。在有风险的人群中，我们的目标是在风险暴露期间预测哪些患者会在定义时间点（t=0）出现某种结局。仅使用包含在那个时刻之前的观察期的患者信息进行预测。

如表 13.1 所示，要定义预测问题，我们必须通过目标队列定义 t=0，从而定义我们想要预测的结果和危险时间。我们将标准预测问题定义如下：

在【目标队列】中，谁将在【风险暴露期间】内产生【结果队列】？

此外，我们需要设计模型，并确定观测数据集以执行内部和外部验证。

选择	描述
目标队列	怎样定义用于预测的目标人群?
结果队列	怎样定义预测结果?
风险暴露期间	用于预测 t=0 相关的时间窗口在哪?
模型	选择哪个算法？算法中包括哪些潜在预测变量?

此框架中的概念适用于所有类型的预测问题，例如：

- 疾病发生和进展
  - **结构**：在新诊断为某*【疾病】*的患者中，谁将在这一*【时间范围内】*出现*【其他疾病或并发症】*？
  - **例子**：在新诊断的房颤患者中，谁将在未来三年内发生缺血性卒中?
- 治疗选择
  - **结构**：在患有某*【适应症】*且接受*【治疗 1】*或者*【治疗 2】*的患者中，哪些患者接受了*【治疗 1】*？
  - **例子**：在接受华法林或利伐沙班的房颤患者中，哪些患者接受了华法林治疗？（比如：倾向模型)
- 治疗反应
  - **结构**：在某*【治疗】*的新使用者中，谁将在特定*【时间窗口】*中经历*【一些效果】*？
  - **例子**：哪些开始服用二甲双胍的患者将连续服用三年?
- 治疗安全
  - **结构**：在某*【治疗】*的新使用者中，谁将在特定*【时间窗口】*中经历*【不良反应】*？
  - **例子**：在华法林的新使用者中，谁会在一年内出现胃肠道出血?

- 治疗依从性
- **结构**：在某*【治疗】*的新使用者中，谁将在特定*【时间窗口】*中达到某一*【依从性指标】*？
- **例子**：在开始服用二甲双胍的糖尿病患者中，哪些人将在一年中>=80%的天数中持续服用？

13.2 数据提取

在创建预测模型时，我们使用监督学习过程（一种机器学习方法），基于一些带有标记的数据示例来推断协变量和结果之间的关系。所以，我们需要从 CDM 中提取目标人群协变量的方法，并获得他们的结果标签。

协变量（又称“预测变量”，“特征”或“独立变量”）描述了患者的特征。协变量可以包括年龄，性别，是否出现某种特定状况和患者记录中药物/操作治疗等。协变量通常由 FeatureExtration 包构造而成，更多细节详见本书第 11 章。在预测问题中我们只能使用患者进入目标队列日期之前或者当天的数据。患者进入目标队列的日期被定义为索引日期。

我们也需要获取所有患者在风险暴露时期内的**结局状态**（也被称为“标签”或“类别”）。如果结局事件在风险暴露期间发生，结局状态则为“阳性”。

13.2.1 数据提取案例

表 13.2 是一个包含 2 个队列的对列表示例。队列定义 ID 1 代表目标队列（比如，“最近被诊断为房颤的患者”）。队列定义 ID 2 代表结果队列（比如，“中风”）。

表 13.2：队列表示例（简洁起见，队列结束日期被省略）

队列定义 ID	主体 ID	队列开始日期
1	1	2000-06-01
1	2	2001-06-01
2	2	2001-07-01

表 13.3 提供了 CONDITION_OCCURENCE(诊断_发生)表的例子。概念 ID320128 代表“原发性高血压”。

表 13.3：CONDITION_OCCURENCE 表示例（简洁起见，此处仅显示 3 列）

个人 ID	状态概念 ID	状态开始日期
1	320128	2000-10-01
2	320128	2001-05-01

基于示例数据，假设风险发生时间为索引日期一年内（即目标队列开始日期），我们可以构建协变量和结局状态。患者 ID 1（状态发生在索引日期后）的协变量“原发性高血压”取值为 0（没有出现），患者 ID 2 的取值为 1（出现）。相似地，患者 ID 1 没有进入结局队列，因此结局状态为 0；患者 ID 2

的结局发生且在索引日期后一年内，因此结局状态为 1。

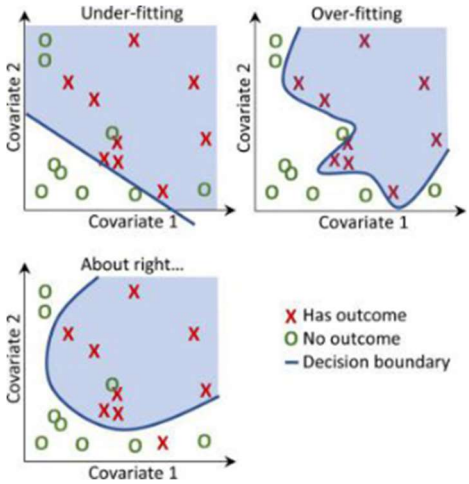
13.2.2 负值（阴性）与数据缺失

观察性医疗健康数据很少显示数值为负值（阴性）或缺失。在前面的例子中，我们仅仅观测到 ID 为 1 的患者在索引日期前没有出现原发性高血压。这有可能是因为当时确实未出现原发性高血压（负值或阴性），也有可能是未被记录（缺失）。重要的是要认识到机器学习算法不能区分负值（阴性）状态和数据缺失，并且只能通过已有数据进行预测。

13.3 模型拟合

在拟合模型时，我们尝试从标记示例中学习协变量和观测到的结局状态之间的关系。假设我们只有两个协变量，收缩压和舒张压，每位患者的数据可以表现为二维空间中的一个点，如图 13.2 所示，点的形状代表患者的结局状态（比如，中风）。

有监督学习模型会力求寻找一个可以将两种状态区分开的最优决策边界。不同的监督学习技术会



形成不同决策边界，其中的一些超变量可以影响决策边界的复杂度。

图 13.2：决策边界

图 13.2 展示了三种不同的决策边界。这些决策边界会被用来推断任何新进入的患者的结局状态，在这里就是一个新的点的结局状态。如果一个新的点落在阴影区域，模型将做出“出现结局”的预测，否则预测为“未出现结局”。理想的决策边界应该将两类数据完美分开。但是，某些情况下，模型会有过度拟合的危险。这将给模型用于未观测过数据上的泛化性带来负面影响。比如，如果数据中可能会包含错误标记或错误定位的数据点，使数据中夹杂噪声，我们并不希望用这些噪声来拟合模型。所以，我们可能会偏向于没有完美区分训练集但是能提炼出“真实”复杂度的决策边界。类似于正则化的技术旨在最小化复杂度的同时将模型表现最大化。

每个监督模型都有其独有的学习决策边界的方法，找出哪个算法最适合你的数据并不容易。正如天下“没有免费的午餐”，没有一个算法可以在所有预测问题中保持最优表现。所以，当开发患者水平的预测模型时，我们推荐尝试带有不同超变量设置的多个监督模型。

PatientLevelPrediction 包中包含如下算法：

13.3.1 正则 Logistic 回归

LASSO（最小绝对值收敛和选择算子、套索算法）logistic 回归属于广义线性模型的范畴，这类模型学习一个变量间的线性模型，最后 logistic 方程将线性组合映射到 0 到 1 间的一个值。基于模型复杂度，正则化 LASSO 训练模型时在目标方程上增加一个额外的成本。这个成本由系数线性组合绝对值的和组成。模型通过最小化成本自动选取特征变量。我们使用 Cyclops 包（logistic、泊松和生存分析的循环坐标下降）执行大规模正则化 logistic 回归。

表 13.4：正则化 logistic 回归的超变量

变量	描述	典型取值
起始方差	先验分布的起始方差	0.1

需要注意的是，模型通过最大化交叉验证中的样本外可能性来优化方差，所有起始方差对最终模型表现的影响很小。但是，选取一个离最优值较远的起始方差会导致拟合时间较长。

13.3.2 梯度提升机

梯度提升机是一项提升集成技术，在我们的框架下包括多个决策树。提升通过迭代加入决策树的方式实现，与此同时，被先前的决策树误判的数据点在训练下一个决策树时被给予更多权重。

表 13.5：梯度提升机中的超变量

变量	描述	典型取值
earlyStopRound	在几轮后停止且不做任何提升	25
learningRate	学习率	0.005，0.01，0.1
maxDepth	决策树中的最大深度层	4，6，17
minRows	节点中的最小数据点	2
ntrees	决策树的个数	100，1000

我们使用的极端梯度提升是梯度提升框架中的高效实现，CRAN 中的 xgboost R 包可用来实现它。

13.3.3 随机森林

随机森林是一种包括多个决策树的装袋集成技术。装袋法将弱分类模型整合成强分类模型，以此降低过度拟合的可能性。为了实现以上效果，随机森林训练多个决策树，每个决策树仅适用一部分变量且每个决策树使用的变量不同。我们的包使用 Python 中的 sklearn RandomForest 来实现此模型。

表 13.6：随机森林中的超变量

变量	描述	典型取值
maxDepth	决策树中的最大深度层	4，10，17
metries	每棵决策树中特征变量的个数	-1=总特征变量的平方根，5，20

ntrees	决策树个数	500

13.3.4 K 临近算法

K 临近算法（KNN）是一种使用距离度量来寻找一个新的未标记数据点最近 K 个标记数据点的算法。新数据点的预测由 K 个最近标记数据点中最普遍的类来决定。KNN 有一个限制：由于模型需要有标记的数据来预测新点，所以通常无法跨数据站点来共享数据。本书包含 OHDSI 开发的 BigKun 包，其可用来实现大规模 KNN 分类模型。

表 13.7：K 临近算法中的超变量

变量	描述	典型取值
k	临近点的个数	1000

13.3.5 朴素贝叶斯

朴素贝叶斯算法在朴素假设 (即每对特征变量在给定分类变量数值的基础上条件独立) 的基础上运用贝叶斯定理。此算法在先验分布和数据从属某一类可能性的基础上计算后验分布。朴素贝叶斯没有超变量。

13.3.6 自适应增强

自适应增强是一项提升集成技术。提升通过迭代的方式加入分类模型实现，与此同时，被先前的分类模型误判的数据点在训练下一个分类模型时被给予更多权重。我们使用 Python 中的 sklearn AdaboostClassifier 来实现此技术。

表 13.8：自适应增强中的超变量

变量	描述	典型取值
nEstimators	停止提升时估计量的最高数量	4
learningRate	学习率将每个分类模型的贡献缩小了 learning_rate 的值。需要权衡 learningRate 和 nEstimators。	1

13.3.7 决策树

决策树是一种分类算法，它使用被选择的个体测试数据遍历能够将变量空间分割的可能性。它旨在寻找具有最高信息增益的分区方式。通过增加区域个数（树的深度），决策树很容易过度拟合，所以通常都需要进行正则化（比如，修建或指定控制模型复杂度的超变量）。我们使用 Python 中的 sklearn DecisionTreeClaasifier 来实现决策树。

表：决策树中的超变量

变量	描述	典型取值

|类权重|“平衡”或“无”|无||maxDepth|决策树的最大深度|10||minImpuritySplit|决策树提早停止生长的阈值。如果节点杂志高于阈值，其将分裂，否则其将称为叶子|10^-7|minSampleLeaf|每片叶子中的最少样本量|10||minSampleSplit|每个分裂中的最少样本量|2|。

13.3.8 多层感知器

多层感知器是使用非线性方程将输入加以权重且包含多层节点的神经网络。第一层是输入层，最后一层是输出层，中间是隐藏层。神经网络通常由反向传播算法训练得出，也就是说，训练集作为输入通过网络结构向前推进直到生成输出，输出和结局状态间的误差通过网络结构被反向传播同时被用来更新线性方程的权重。

表 13.9： 多层感知器中的超变量

变量	描述	典型取值
Alpha	L2 正则化	0.00001
size	隐藏节点的个数	4

13.3.9 深度学习

包含深度网络、卷积神经网络和递归神经网络在内的深度学习与多层感知器相似，但是深度学习算法包含了多层隐藏层，其被用来学习有助于预测的潜在表达。在 PatientLevelPrediction 包中有一个独立的专门的段落对这些模型和超变量进行了更详尽的描述。

13.3.10 其他算法

也可以通过增加其他算法来进行患者水平的预测，但是这不在本章的讨论范围内。读者可以在 PatientLevelPrediction 包“增加自定义的患者水平预测算法”的段落中找到更多详细说明。

13.4 评估预测模型

13.4.1 评估类型

通过对比模型预测结果与实际观测结果的一致性，评估预测模型。

预测模型的构建与验证应使用不同的数据集，否则会造成模型过度拟合（见 13.3），并且可能对新患者的预测效果不佳。

内部验证与外部验证的区别：

内部验证：使用从同一数据库中提取的不同数据集来预测和评估模型。

外部验证：在一个数据库中开发模型，并在另一个数据库中进行评估。

内部验证有两种方法：

保留集法：是将一个数据库拆分为两个独立的数据集，即一个训练集和一个测试集（保留集）。训练集用于学习模型，测试集用于评估模型。我们可以简单地将患者随机分为训练和测试集，或者随机选择：

-根据时间（时间验证）分割数据，例如在特定日期之前对数据进行训练，然后在该日期之后对数据进行评估。这可能会告诉我们，我们的模型是否适用于不同的时间段。

-根据地理位置分割数据（空间验证）。

交叉验证法：当数据有限时，可采用交叉验证。数据被分成 n 个相等的集合，其中 n 需要预先指定（例如 n=10）。对于每一个集合，一个模型都对除该集合中的数据之外的所有数据进行训练，并用于保留集的预测。因此，该数据库中的所有数据都被用于评估模型的建立。在患者水平预测框架中，可使用交叉验证来选择最佳超变量。

外部验证旨在评估预测模型用于开发环境之外数据的预测能力，预测模型的这种可移植性能力是很重要的。不同的数据库可能代表不同的患者群体、不同的医疗系统和不同的数据获取过程。我们认为，在大型数据库中使用预测模型进行外部验证，是可以让临床实践接受和实施的关键步骤。

13.4.2 模型评估指标

阈值评估

预测模型为每个患者分配一个介于 0 和 1 之间的值，该值对应于患者在风险时间内获得结果的风险。值为 0 表示 0%风险，值为 0.5 表示 50%风险，值为 1 表示 100%风险。常见的指标，如准确性、敏感性、特异性、阳性预测值，可以通过首先指定一个阈值来计算，用该阈值将患者分类为在风险期内是否有结果。例如，如表 13.10 所示，如果我们将阈值设置为 0.5，则患者 1、3、7 和 10 的预测风险大于或等于阈值 0.5，因此他们将被预测为有结果。所有其他患者的预测风险低于 0.5，因此他们将被预测为没有结果。

如果预测模型结果与风险期内的实际结果均为阳性，称为真阳性（TP）。如果预测模型结果为阳性，而风险期内的实际结果均为阴性，称为假阳性（FP）。如果预测模型结果与风险期内的实际结果均为阴性，称为真阴性（TN）。如果预测模型结果为阴性，而风险期内的实际结果均为阳性，称为假阴性（FN）。

Table 13.10: 使用预测概率阈值的示例

Patient ID	Predicted risk	Predicted class at 0.5 threshold	Has outcome during time-at-risk	Type
1	0.8	1	1	TP
2	0.1	0	0	TN
3	0.7	1	0	FP
4	0	0	0	TN
5	0.05	0	0	TN
6	0.1	0	0	TN
7	0.9	1	1	TP
8	0.2	0	1	FN
9	0.3	0	0	TN
10	0.5	1	0	FP

可以计算以下基于阈值的指标：

准确性: (T P + T N)/(T P + T N + F P + F N)

敏感性: T P/(T P + F N)

特异性： T N/(T N + F P)

阳性预测值: T P/(T P + F P)

请注意，如果降低阈值，这些值可以减小或增大。降低阈值可以增加分母。如果之前设置的阈值过高，新的结果可能都是真阳性，这将增加阳性预测值。如果之前的阈值刚好或过低，进一步降低阈值将引入假阳性，降低阳性预测值。对于灵敏度，分母不依赖于阈值（T P+F N 是常数）。这意味着降低阈值可以通过增加真实阳性结果的数量来提高灵敏度。降低阈值也可能使灵敏度保持不变，而阳性预测值则会波动。

判别能力

判别能力是指将更高的风险分配给在风险期内会发生结局的人。受试者操作 m 特征曲线（ROC）是通过绘制 1-特异性（X 轴）和在所有可能的阈值上的灵敏度（Y 轴）来创建的。本章下文给出 ROC 图示例（见图 13.17）。ROC 下面积（AUC）代表总体诊断能力，若 AUC 为 0.5，则说明预测是随机分配风险，AUC 为 1 表示模型能够完全预测正确。大多数已发表的预测模型的 AUC 在 0.6-0.8 之间。AUC 可用来确定在风险期内发生结局的患者与没有发生结局的患者之间预测风险分布的差异。如果 AUC 很高，那么分布将大部分分离，而当有很多重叠时，AUC 将接近 0.5，如图 13.3 所示。

对于罕见的结果，即使是具有高 AUC 的模型也可能不适用，因为对于高于给定阈值的每一个阳性值，也可能有许多阴性值（即阳性预测值将很低）。当干预措施会造成严重的健康风险或干预成本较高（健康风险和/或金钱）时，高假阳性率是不希望看到的结果。当结果罕见或很少见时，建议使用另一种称为精确召回率曲线下面积 (AUPRC) 的测量方法。AUPRC 是以灵敏度 (也称为召回率) 为横坐标，正预测值为纵坐标生成的曲线下面积。

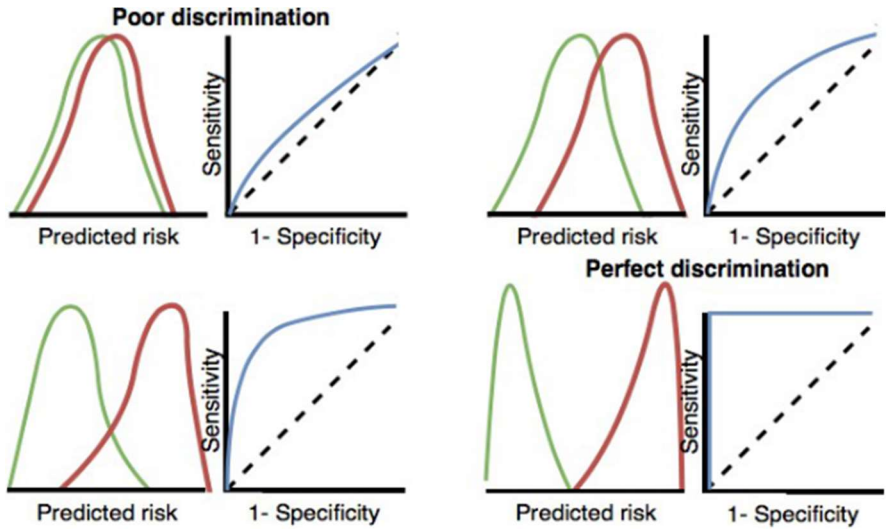


图 13.3：ROC 图与判别能力的关系。如果两类预测风险分布相似，则 ROC 接近对角线，AUC 接近 0.5。

校准

校准是指模型分配正确风险的能力。例如，如果模型给 100 名患者分配了 10%的风险，那么 10 名患者应该在有风险的时间发生结局。如果模型给 100 名患者分配了 80%的风险，那么 80 名患者应

该在风险期内发生结局。校准通常是根据预测的风险将患者按十分位数分配并进行计算，在每组中计算平均预测风险和在风险时间内发生结局的患者比例。然后，我们绘制这十个点（在 Y 轴上预测风险，并观察 X 轴上的风险），并看看它们是否落在 x= y 线上，表明模型校准良好。本章后文给出了校准图的示例（图 13.18）。我们还利用这些点拟合了一个线性模型来计算截距和梯度。如果梯度大于 1，则模型分配的风险高于真实风险；如果梯度小于 1，则模型分配的风险低于真实风险。另外，我们还在 R 包中实现了平滑校准曲线，以便更容易发现预测风险和观测风险之间的非线性关系。

13.5 设计患者水平预测研究

在这部分我们将展示如何设计预测研究。第一步是清晰地定义预测问题。有趣的是，许多文章发表中，预测问题的定义并不明确，比如索引日期（目标队列的开始日期）的定义。其他人无法对一个定义模糊的预测问题进行外部验证，更不用说在临床实践中实施了。在患者水平预测的框架下，我们要求明确预测问题，也就是说表 13.1 中的关键选项需要定义清楚。这里我们将以“预测安全性”类预测问题为例并逐步展示这一过程。

13.5.1 问题定义

血管性水肿是 ACE 抑制剂众所周知的副作用，ACE 抑制剂标签中报道的血管性水肿发生率为 0.1% 至 0.7%（Byrd et al.,2006）。 监测患者的这一不良反应很重要，尽管血管性水肿很罕见，但其可能危及生命、导致呼吸骤停和死亡（Norman et al.,2013）。此外，如果最初没有认识到血管性水肿的可能性，发现这一原因需要一系列昂贵的检查。（Norman et al.，2013; Thompson and Frable，1993）。除了非裔美国患者有较高风险外，没有发现其他诱发血管紧张素转换酶抑制剂相关血管性水肿发展的因素。 （Byrd 等，2006）这一不良反应大多发生在初始治疗的第一周或一个月内，通常在接受初始剂量的几小时内（Cicardi et al.，2004）。但是，也有些病例可能发生在开始治疗数年后。 （O’Mara 和 O’Mara，1996）目前尚无专门用于识别该风险的诊断检测方法。如果我们能够识别出那些有危险的患者，医生就可以采取措施，例如放弃 ACE 抑制剂，改用另一种高血压药物。

我们将把患者水平的预测框架应用于观察性医疗数据，以解决以下患者水平的预测问题

在刚开始使用 ACE 抑制剂的患者中，谁会在下一年中出现血管性水肿?

13.5.2 研究人群定义

我们最终用来开发模型的研究人群通常是目标队列的一个子集，因为我们可能根据结局事件相关标准对目标队列进行筛选，或者我们想要做针对目标队列亚群的敏感性分析。为此，我们必须回答以下问题：

在目标队列开始之前，我们需要的最少观察时间是多长? 该选择可能取决于训练集中可用的患者时间数据，以及我们将来拟运用该模型的数据源环境中的可用时间。最短观察时间越长，可用于患者特征提取的基线历史时间越长，但是进入分析的患者就越少。此外，临床原因可能导致较短或较长的回溯期。在我们的例子中，我们将使用 365 天以前的历史记录作为回顾期（清除期）。

患者能否多次进入目标队列? 在目标队列的定义中，一个人在不同的时间阶段可能多次进入队列，比如，如果他们经历不同的疾病发作期或者在不同时期内接受过医疗产品的处置。队列定义不一定限定患者只进入一次，但根据特定的患者水平预测的问题，我们希望将队列限定在第一个符合条件的疾病发

作期。在我们的例子中，患者只能进入目标队列一次，因为我们需要研究患者第一次使用 ACE 抑制剂的数据。

如果患者在之前经历过结局，是否允许他们进入队列？ 如果患者在进入目标队列前经历了结局，我们是否将他们加入到目标队列？这取决于特定的预测问题，如果我们想要预测结局事件的首次发生，已经发生过结局事件的患者不再有首次发生的风险，所以这部分患者会被目标队列排除。在其他情况下，我们可能想要预测结局的普遍发生，已经发生过结局事件的患者可以被包含在分析中，并且已发生的结局本身可以用来预测未来的结果。在我们的例子中，以前发生过血管性水肿的患者不会被包含在分析中。

相对于目标队列开始时间，如何定义预测时间窗？ 回答这个问题，需要做出以下两个决定。第一，风险暴露时间窗是否与目标队列窗同时开始或者发生在其后？使风险暴露时间窗的开始发生在目标队列窗后的考虑包括：我们想要避免较晚进入记录但是实际上发生在目标队列开始之前的结果，或者我们想为预防结果的干预留下理论上可行的时间差。第二，我们需要设定时间窗终点来定义风险暴露时间区间(time-at-risk)，以此作为相对于目标队列开始或结束日期的天数抵消。在我们的问题中，我们将风险暴露时间窗定义为目标队列开始后的 1-365 天。

是否要求最短风险暴露时间区间？ 我们需要决定是否将在风险暴露时间窗结束前未出现结局就提前离开队列的患者包含到最终的分析中。这部分患者可能在未继续观察的时间窗以外的时间里出现结局。我们的预测问题中，回答为“是”，我们要求一个最短风险暴露期间。此外，我们需要决定这一限制是否同样适用于出现了结局的患者，或者所有出现该结局的患者都会被包含，无论总体风险暴露时间有多长。比如，如果结局是死亡，经历这一结局的患者有可能在整个风险时间窗结束前就会被审查。

13.5.3 模型开发设置

要开发预测模型，必须确定希望训练哪种算法。我们认为针对某个预测问题的最佳算法的选择是经验问题，也就是说，我们倾向于让数据说话，尝试不一样的方法从而发现最佳算法。在我们的框架中，我们尝试了 13.3 中的许多算法，并且支持尝试其他算法。在目前的例子中，为简单起见，我们只挑选了一个算法：梯度提升机。

此外，我们需要挑选训练模型的协变量。在我们的例子中，性别，年龄，所有状况，药物和药物组，就诊次数等被设为协变量。我们将在索引日期之前的一年中以及索引日期之前的任何时间查找这些临床事件。

13.5.4 模型评价

最后，我们需要定义模型的评价方式。简洁起见，这里我们选择内部验证法。此法需要定义训练集，测试集以及将患者安排到每个集合的方式。这里我们使用典型的 75%-25%分割。需要注意的是，在大型数据集中我们可以用更多的数据来训练模型。

13.5.5 研究总结

表 13.11 完整地定义了我们的研究。

表 13.11：主要设计选项示例

选项	取值

目标队列	第一次使用 ACE 抑制剂的患者。观测时间短于 365 天，或前期已发生过血管性水肿的患者将被排除。
结局队列	血管性水肿
风险暴露期间	队列开始后的 1-365 天。我们要求最少有 364 天的风险期。
模型	梯度提升机，超变量定义如下： ntree: 5000, max depth: 4 或 7 或 10，learning rate: 0.001 或 0.01 或 0.1 或 0.9。协变量包括性别，年龄，状态，药物，药物组和就诊次数。数据分割：75%训练集-25%测试集，患者随机分配。

13.6 在 ATLAS 中的操作

点击左侧 ATLAS 菜单中的 Prediction 按钮，可以打开预测研究设计界面。创建新的预测研究，并对这项研究取一个容易识别的命名。点击 按钮可随时保存研究设计。

在预测设计功能中，有 4 个部分：预测问题设置、分析设置、执行设置和培训设置。下文将逐一讨论：

13.6.1 预测问题设置

我们选择目标人群队列和结局队列进行分析。将为目标人群队列和结局队列所有组合开发一个预测模型。例如，如果我们指定两个目标人群和两个结局，就相当于我们设定了四个预测问题。为了选择目标人群，我们需要在 ATLAS 中预先定义它。第 10 章描述了如何建立队列。附录提供了本例中使用的目标（附录 B.1）和结局（附录 B.4）队列的完整定义。要将目标人群添加到队列中，可点击“Add Target Cohort”按钮。通过点击“Add Outcome Cohort”也可以添加。完成后，对话框应该如图 13.4 所示。

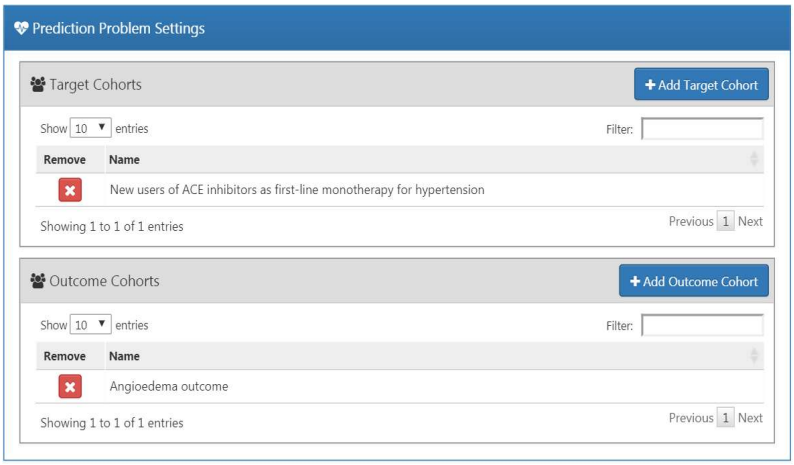


图 13.4：预测问题设置

13.6.2 分析设置

分析设置允许选择监督学习算法、协变量、和人群。

模型设置

我们可以选择一个或多个监督学习算法进行模型开发。可点击“Add Model Settings”按钮，下拉菜单将显示包含 ATLAS 界面中当前支持的所有模型。通过点击下拉菜单中的名称，可以选择要包含在研究中的监督学习模型。这将显示该特定模型的视图，允许选择超变量值。如果提供多个值，则对所有可能的值组合执行网格搜索，以使用交叉验证选择最佳组合。在本章例子中，选择了 gradient boosting machines（GBM），并设置了图 13.5 中指定的超变量。

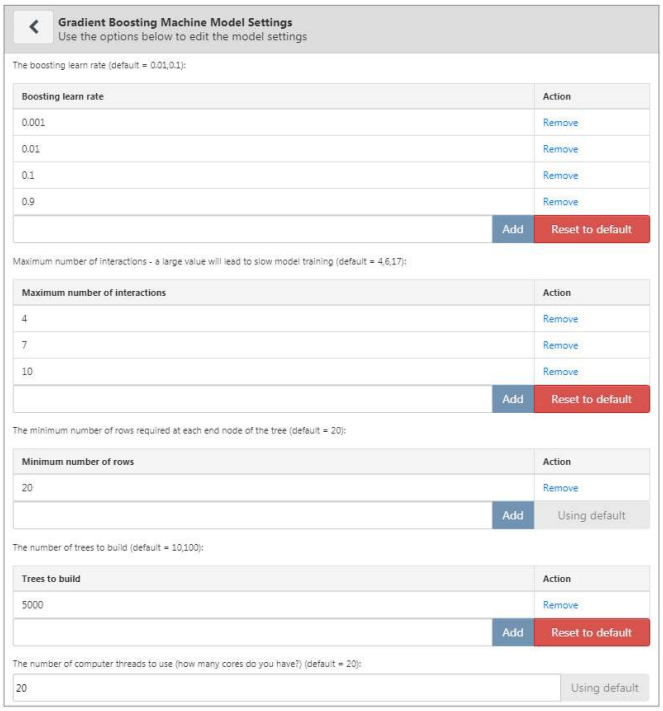


图 13.5：Gradient Boosting Machine 设置

协变量设置

可以从 CDM 格式的观察数据中提取一组标准协变量，在协变量设置视图中，可以选择要包含的标准协变量。我们可以定义不同类型的协变量设置，并且每个模型将使用其指定的协变量设置分别创建。要在研究中添加共变设置，请单击“Add Covariate Settings”，将打开协变量设置视图。

协变量设置视图的第一部分是 exclude/include 选项。任何概念都可以构造协变量。然而，我们希望纳入或排除特定的概念，例如，关联目标群组的定义的概念。若想包含某些特定概念，请在 ATLAS 中创建一个概念集，然后在 **“What concepts do you want to include in baseline covariates in the patient-level prediction model? (Leave blank if you want to include everything) ”** 通过点击  选择概念集。我们可以通过对 **“Should descendant concepts be added to the list of included concepts?”** 的问题回答“是”， 自动将所有子概念添加到概念集。 同样的过程可以重复 **“What concepts do you want to exclude in baseline covariates in the patient-level prediction model? (Leave blank if you want to include everything)”**，允许删除与所选概念对应的协变量。最后一个选项 **“A comma delimited list of covariate IDs that should be**

restricted to” 允许我们添加一组以逗号分隔的协变量 ID（而不是概念 ID），这些 ID 将仅包含在模型中。此选项仅适用于高级用户。完成后，包含和排除设置应该如图 13.6 所示。

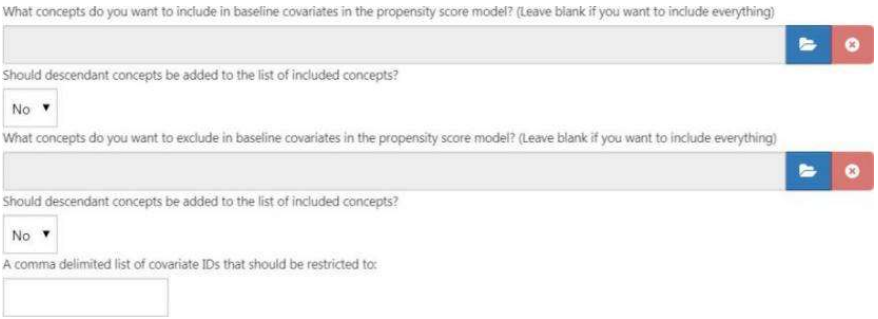


图 13.6：协变量纳入排除设置

性别：表示男性或女性性别的二元变量

年龄：表示年龄的连续变量

年龄组：每 5 岁一个二元变量（0-4、5-9、10-14、…、95+）

种族：每个种族一个二元变量，1 表示患者有该种族的记录，否则为 0

人种：每个种族一个二元变量种族，1 表示患者记录了该种族，否则为 0。

开始年份：每个队列开始年份的二元变量，1 表示患者队列有开始年份，否则为 0。由于我们希望将模型应用于未来，因此将开始年份包括在内通常是没有意义的。

开始月份：每个队列开始日期月的二元变量，1 表示患者的队列有开始月份，否则为 0。

观察时间之前：[不建议用于预测]与患者在队列开始日期之前在数据库中的天数相对应的连续变量

观察时间之后：[不建议用于预测]与患者在数据库中的天数相对应的连续变量队列开始日期

队列中时间：一个连续变量，对应于患者在队列中的天数（队列结束日期减去队列开始日期）

开始年份和月份：[不建议用于预测]每个队列起始日期年月组合的二元变量，1 表示患者队列有起始年月，否则为 0。

完成后，该部分应如图 13.7 所示。

图 13.7：协变量选择

标准协变量提供三个灵活的时间间隔：

结束时间：相对于队列起始日期结束的时间间隔[默认为 0]

长期：[默认- 365 天到队列开始日期之前的末日]

中期：[默认- 180 天到队列开始日期之前的末日]

短期：[默认- 队列开始日期前 30 天结束]

完成后，该部分应如图 13.8 所示。

Time bound covariates

Set the time windows for the time bound covariates in days relative to the cohort index

	Any Time Prior	Long Term	Medium Term	Short Term	End Days
Time Windows	All Time	-365	-180	-30	0

图 13.8：时间相关协变量选择

下一步的选项是从 era 表中提取协变量：

状况（疾病）：为每个选择的疾病概念 ID 和时间间隔构建协变量，如果患者的概念 ID 带有 era（即疾病在时间间隔内开始或结束，或在时间间隔前后开始和结束）在疾病 era 表中队列开始日期之前的指定时间间隔内，协变量值为 1，否则为 0。

状况（疾病）组：为选择的每个疾病概念 ID 和时间间隔构建协变量，如果患者在疾病 era 表中队列开始日期之前的指定时间间隔内具有概念 ID **或任何带有 era 的子概念 ID**，则协变量值为 1，否则为 0。

药物：为选择的每个药物概念 ID 和时间间隔构建协变量，如果患者在药物 era 表中队列开始日期之前的指定时间间隔内具有带有 era 的概念 ID，则协变量值为 1，否则为 0。

药物组：为选择的每个药物概念 ID 和时间间隔构建协变量，如果患者在药物 era 表中队列开始日期之前的指定时间间隔内具有概念 ID **或具有 era 的任何子概念 ID**，则协变量值为 1，否则为 0。

重叠时间间隔设置意味着药物或条件 era 应在队列开始日期之前开始，并在队列开始日期之后结束，因此它与队列开始日期重叠。**era start** 选项限制查找在所选时间间隔内开始的条件或药物 era。

完成后，该部分应如图 13.9 所示。

Set the time bound era covariates

Domain	Any Time Prior	Long Term (-365 days)	Medium Term (-180 days)	Short Term (-30 days)	Overlapping	Era Start: Long Term (-365 days)	Era Start: Medium Term (-180 days)	Era Start: Short Term (-30 days)
Condition	☐	☐	☐	☐	☐	☐	☐	☐
Condition Group	☐	☑	☐	☑	☐	☐	☐	☐
Drug	☐	☐	☐	☐	☐	☐	☐	☐
Drug Group	☐	☑	☐	☑	☐	☐	☐	☐

图 13.9: 有时间限制的 era 协变量。

下一步选项为不同的时间间隔选择对应于每个域中的概念 ID 的协变量：

疾病状态：为每个选择的疾病状态概念 ID 和时间间隔构造协变量，如果患者在疾病发生表中的队列开始日期之前的指定时间间隔内记录了概念 ID，协变量值为 1，否则为 0。

住院患者主要疾病状态：针对住院患者拟观察疾病状况对应的主诊断，在 condition_occurrence 表中，设置相应的二元协变量。

药物：为选择的每个药物概念 ID 和时间间隔构建协变量，如果患者在药物暴露表中的队列开始日期之前的指定时间间隔内记录了概念 ID，则协变量值为 1，否则为 0。

操作：为选择的每个操作概念 ID 和时间间隔构建协变量，如果患者在操作发生表中队列开始日期之前的指定时间间隔内记录了概念 ID，则协变量值为 1，否则为 0。

检验：为选择的每个检验概念 ID 和时间间隔构建协变量，如果患者在检验表中队列开始日期之前

的指定时间间隔内记录了概念 ID，则协变量值为 1，否则为 0。

检验值：为每个检验概念 ID 构建协变量，选择一个值和时间间隔，如果患者在检验表中队列开始日期之前的指定时间间隔内记录了概念 ID，则协变量值为测量值，否则为 0。

检验范围：指示测量值是否低于、在正常范围内或高于正常范围的二元协变量。

观察：为每个观察概念 ID 和时间间隔构建协变量选择，如果患者在观察表中队列开始日期之前的指定时间间隔内记录了概念 ID，则协变量值为 1，否则为 0。

器械：为选择的每个器械概念 ID 和时间间隔构建协变量，如果患者在器械表中队列开始日期之前的指定时间间隔内记录了概念 ID，则协变量值为 1，否则为 0。

就诊次数：为每次就诊和选定的时间间隔构建协变量，并将该时间间隔内记录的访视次数作为协变量值进行计数。

就诊概念计数：为选择的每次就诊、域和时间间隔构建协变量，并将就诊类型和时间间隔期间记录的每个域的记录数作为协变量值。distinct count 选项统计每个域和时间间隔的 distinct concept ID 数。

完成后，该部分应如图 13.10 所示。

Set the time bound covariates

Domain	Any Time Prior	Long Term (-365 days)	Medium Term (-180 days)	Short Term (-30 days)	Distinct Count		
					Long Term (-365 days)	Medium Term (-180 days)	Short Term (-30 days)
Condition	☐	☐	☐	☐	☐	☐	☐
Condition - Primary Inpatient	☐	☐	☐	☐			
Drug	☐	☐	☐	☐	☐	☐	☐
Procedure	☐	☑	☐	☐	☐	☐	☐
Measurement	☐	☑	☐	☐	☐	☐	☐
Measurement - Value	☐	☐	☐	☐			
Measurement - Range Group	☐	☑	☐	☐			
Observation	☐	☑	☐	☐	☐	☐	☐
Device	☐	☑	☐	☐			
Visit - Count		☑	☐	☐			
Visit - Concept Count		☐	☐	☐			

图 13.10：时间设置

最后一个选择是是否将常用的风险得分作为协变量。完成后，风险评分设置应如图 13.11 所示。

Set the index score covariates

Index Score Type	
$CHADS_2$	☐
CHA_2DS_2VASc	☑
DCSI	☑
Charlson	☑

图 13.11：风险协变量设置

人群设置

人群设置是将附加包含条件应用于目标人群的位置，也是定义风险暴露期间（time-at-risk）的位置。若要将人群设置添加到研究中，请单击“Add Population Settings”按钮。这将打开“人群设置”

视图。第一组选项允许用户指定处于风险期的时间。这是我们观察兴趣结果是否发生的时间间隔。如果患者在风险期内有结果，我们将其归类为“Has outcome”，否则归类为“No outcome”。**“Define the time-at-risk window start, relative to target cohort entry:”**定义风险暴露期间的开始，相对于目标队列开始或结束日期。同样，**“Define the time-at-risk window end:”**定义了风险暴露期间的结束。

“Minimum lookback period applied to target cohort” 指定最小基线周期，即在队列开始日期之前连续观察患者的最少天数。默认值为 365 天。扩大最小回顾天数无疑会提供病人更完整的资料（因为他们一定被观察了更长时间），但也将会过滤掉那些不满足最少观察天数的病人。

如果**“Should subjects without time at risk be removed? ”**设置为“是”，则还需要设置**“Minimum time at risk”**的值。这样就能够允许从队列中去除失访的患者，即在风险暴露期离开队列的患者。例如，如果处于风险暴露期的时间是从队列开始的第 1 天到队列开始的第 365 天，那么整个处于危险期的时间间隔是 364 天（365-1）。如果我们只想包括观察整个间隔的患者，那么我们将最小风险时间设置为 364。如果只要观察患者在风险期的前 100 天，那么我们选择的最小风险时间是 100。如果我们将“Should subjects without time at risk be removed?”设置‘不’，那么这将保留所有患者，甚至那些在风险暴露期从队列中退出的患者。

选项**“Include people with outcomes who are not observed for the whole at risk period? ”**与上一个选项有关。如果设置为“是”，则始终保留在风险时间内出现结局的患者，即使他们在指定的最短时间内未被观察到。

选项**“Should only the first exposure per subject be included? ”**只有当我们的目标队列中存在同一个患者多次暴露，即具有不同队列开始日期的患者时才有用。在这种情况下，选择“是”将导致在分析中仅保留每个患者的最早目标队列日期。否则，患者可以多次出现在数据集中。

设置“**Remove patients who have observed the outcome prior to cohort entry?** ”到“是”将删除在风险开始日期之前有结局事件的患者，因此模型适用于以前从未经历过结局的患者。如果选择“否”，则患者可能之前有过结局事件。通常情况下，事先有结局是非常有预测性的，在有风险的时候会有结局。

完成后，“人群设置”对话框应如图 13.12 所示。

图 13.12：人群设置

现在我们已经完成了分析设置，整个对话框应该如图 13.13 所示。

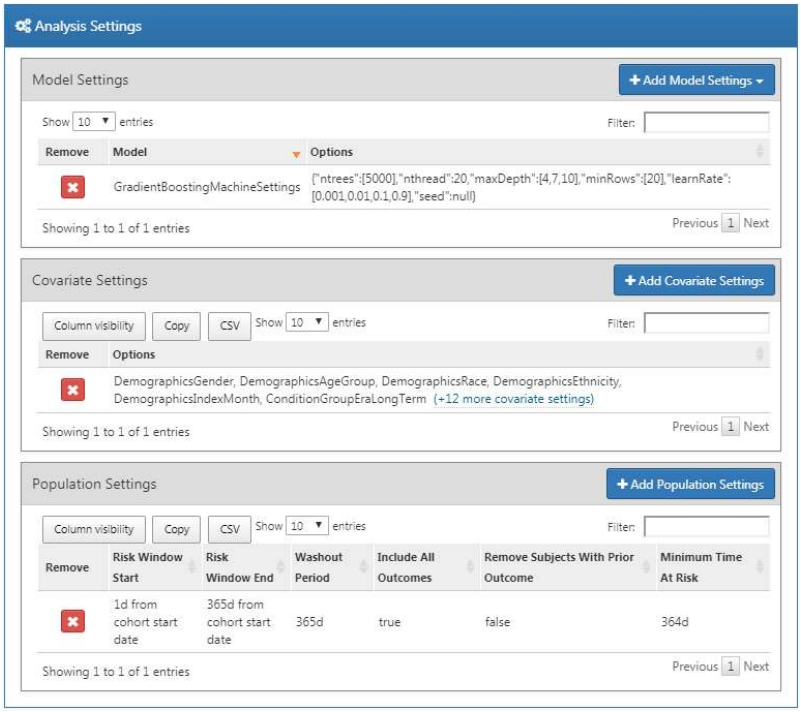


图 13.13：分析设置

在本例子中，我们选择的设置如图 13.14 所示。

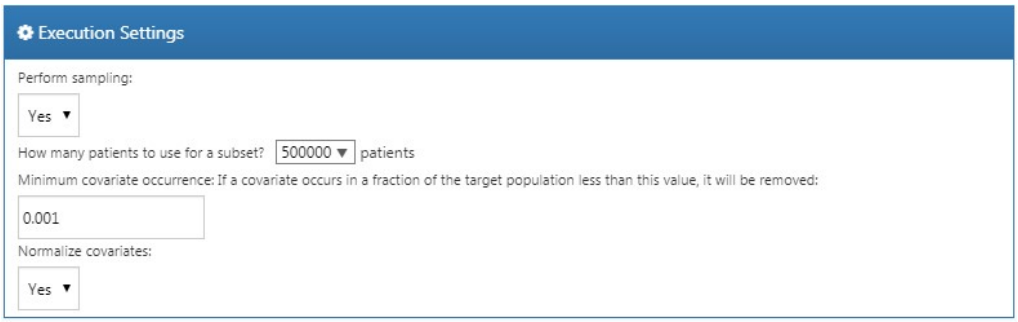


图 13.14：执行设置

13.6.4 训练设置

有四个选项：

•“ **Specify how to split the test/train set** ”： 选择是按人员（按结果分层）还是按时间（较旧数据用于训练模型，较新数据用于评估模型）来划分训练/测试数据。

•“ **Percentage of the data to be used as the test set (0-100%)**”： 选择一定比例的数据作为测试集（默认比例为 25%）。

•“ **The number of folds used in the cross validation**” ： 通过不同交叉验证的折叠数，来选择最优的超参数（默认值为=3）。

• **“The seed used to split the test/train set when using a person type test Split (optional)”** ：当选择按人员类型划分测试集时，一般用随机种子划分训练/测试集。

在本例子中，我们选择的设置如图 13.15 所示。

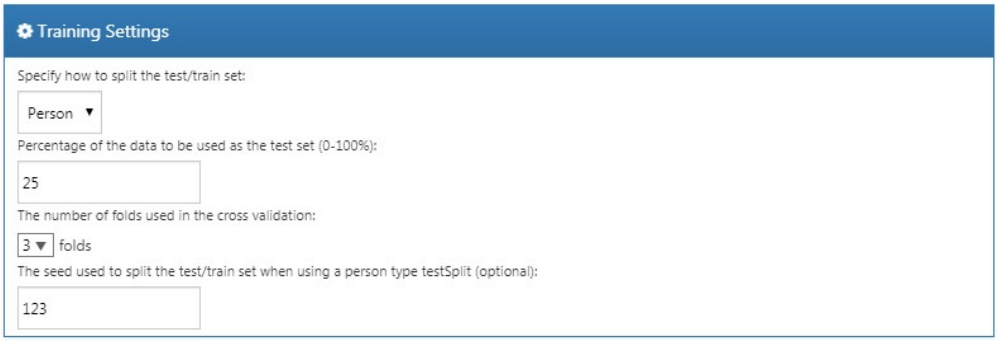


图 13.15：训练设置

13.6.5 导入和导出一项研究

若要导出一项研究，请单击“Utilities”下的“Export”选项卡。ATLAS 将生成可直接复制粘贴到文件的 JSON 文本，包含运行此研究所需的所有数据，如研究名称、队列定义、所选模型、协变量、设置等。要导入一项研究，请单击“Utilities”下的“Import”选项卡。将患者水平预测研究的 JSON 内容粘贴到此窗口中，然后单击其他选项卡按钮下的导入按钮。请注意，这将覆盖当前研究的所有现有设置，因此通常使用新的空白研究设计来完成此操作。

13.6.6 下载研究包文件

单击“Utilities”下的“Review & Download”选项卡。在“Download Study Package”部分，输入 R 包的描述性名称，注意，ATLAS 会自动删除 R 中文件名里的任何非法字符。单击 

可将 R 包下载到本地文件夹。

13.6.7 运行研究

运行 R 包需要按照第 8.4.5 节中的说明安装 R、RStudio 和 Java。还需要 PatientLevelPrediction 包，它可以使用以下命令安装在 R 中：

```
install.packages("drat")
drat::addRepo("OHDSI")
install.packages("PatientLevelPrediction")
```

某些机器学习算法需要安装额外的软件。有关如何安装 PatientLevelPrediction 软件包的完整说明，请参阅 “Patient-Level Prediction Installation Guide”。要使用研究的 R 包，我们建议使用 R Studio。如果在本地运行 R Studio，请解压缩 ATLAS 生成的文件，然后双击 Rproj 后缀文件在 R Studio 中打开它。如果在 R Studio Server 上运行 R Studio，请单击 

.上传和解压缩文件，然后单击 Rproj 后缀文件以打开项目。在 R Studio 中打开项目后，可以打开 README 说明文件，并按照说

明进行操作。请确保将所有文件路径更改为当前系统上已有的路径。

13.7 在 R 中进行研究

使用 ATLAS 实现研究设计的另一种方法是用 R 编写研究代码。我们可以使用 PatientLevelPrediction 包中提供的功能。该软件包使用数据库中已转换为 OMOP CDM 的数据实现数据提取、模型构建和模型评估。

13.7.1 队列初始化

我们首先需要初始化目标队列和结果队列。第 10 章描述了如何初始化队列。附录提供了目标队列（附录 B.1）和结果队列（附录 B.4）的完整定义。在这个例子中，我们假设 ACE 抑制剂组 ID 为 1，血管性水肿组 ID 为 2。

13.7.2 数据提取

首先需要告诉 R 如何连接到服务器。PatientLevelPrediction 使用 DatabaseConnector 包，提供了一个名为 CreateConnectionDetails 的函数。输入*? createConnectionDetails* 可查看适用于各种数据库管理系统（DBMS）的特定设置。例如，可以使用以下代码连接到 PostgreSQL 数据库：

```
library(PatientLevelPrediction)
 connDetails <- createConnectionDetails(dbms = "postgresql",
                                        server = "localhost/ohdsi",
                                        user = "joe",
                                        password = "supersecret")

 cdmDbSchema <- "my_cdm_data" cohortsDbSchema <- "scratch" cohortsDbTable
 <- "my_cohorts" cdmVersion <- "5"
```

最后四行代码定义了 cdmDbSchema、cohortsDbSchema 和 cohortsDbTable 变量，以及 CDM 版本。接下来我们将使用这些变量来告诉 R CDM 格式的数据存储在何处，已创建的感兴趣的队列在何处，以及 CDM 使用的版本。请注意，对于 Microsoft SQL Server，数据库架构需要同时指定数据库和架构，例如 cdmDbSchema <- “my_cdm_data.dbo”。

首先，通过计算队列条目的数量来验证队列创建是否成功是有意义的：

```
sql <- paste("SELECT cohort_definition_id, COUNT(*) AS count", "FROM
@cohortsDbSchema.cohortsDbTable",
"GROUP BY cohort_definition_id") conn <- connect(connDetails)
renderTranslateQuerySql(connection = conn,
                        sql = sql,
                        cohortsDbSchema = cohortsDbSchema,
                        cohortsDbTable = cohortsDbTable)
```

```
##  cohort_definition_id   count
## 1                    1   527616
## 2                    2   3201
```

我们可以通过 PatientLevelPrediction 提取所有必要的数据进行分析。使用 FeatureExtraction 包提取协变量。有关 FeatureExtraction 包的更多详细信息，请参阅说明。在我们的示例研究中，我们决定使用以下设置：

```
covariateSettings <- createCovariateSettings( useDemographicsGender = TRUE,
                          useDemographicsAge = TRUE,
                          useConditionGroupEraLongTerm = TRUE,
                          useConditionGroupEraAnyTimePrior = TRUE,
                          useDrugGroupEraLongTerm = TRUE,
                          useDrugGroupEraAnyTimePrior =TRUE,
                          useVisitConceptCountLongTerm = TRUE,
                          longTermStartDays =-365,
                          endDays = -1)
```

提取数据的最后一步是运行 getPlpData 函数并输入连接详细信息、存储队列的数据库架构、队列和结果的定义 ID，以及洗脱期（即在队列索引日期之前，能在数据中观察到的天数最小值），最后输入先前构建的协变量设置。

```
plpData <- getPlpData(connectionDetails = connDetails,
                      cdmDatabaseSchema = cdmDbSchema,
                      cohortDatabaseSchema = cohortsDbSchema,
                      cohortTable = cohortsDbSchema,
                      cohortId = 1,
                      covariateSettings = covariateSettings,
                      outcomeDatabaseSchema = cohortsDbSchema,
                      outcomeTable = cohortsDbSchema,
                      outcomeIds = 2,
                      sampleSize = 10000)
```

getPlpData 函数还有许多其他参数，这些参数都记录在 PatientLevelPrediction 使用手册中。生成的 plpData 对象使用包 ff 来存储信息，以确保即使数据量很大，R 也不会耗尽内存。创建 lpdataobject 可能需要花费大量的计算时间，将其保存以备将来的会话使用可能是一个好主意。plpData 使用 ff，因此我们将无法使用 R 的常规保存函数，而是使用 savePlpData 函数：

```
savePlpData(plpData, "angio_in_ace_data")
```

我们可以使用 loadPlpData()函数在之后的会话中加载数据。

13.7.3 附加入选标准

通过对两个早期定义的队列应用附加约束来获得最终研究人群，例如，可以强化最小风险时间约束（requireTimeAtRisk， minTimeAtRisk），也可以指定此约束是否也适用于结果队列的患者

（includeAlloutcomes）。现在我们指定相对于目标队列开始的风险窗口的开始和结束。例如，如果我们希望风险窗口，从风险队列开始 30 天后开始，并在一年之后结束，我们可以设置 riskWindowStart=30 和 riskWindowEnd=365。在某些情况下，风险窗口需要从队列结束日期开始。这可以通过设置 addExposureToStart=TRUE 来实现，它会将队列（暴露）时间添加到开始日期。

下面的示例，是我们在研究中采用的所有的设置：

```
population <- createStudyPopulation(plpData = plpData,
                                     outcomeId = 2,
                                     washoutPeriod = 364,
                                     firstExposureOnly = FALSE,
                                     removeSubjectsWithPriorOutcome = TRUE,
                                     priorOutcomeLookback = 9999,
                                     riskWindowStart = 1,
                                     riskWindowEnd = 365,
                                     addExposureDaysToStart = FALSE,
                                     addExposureDaysToEnd = FALSE,
                                     minTimeAtRisk = 364,
                                     requireTimeAtRisk = TRUE,
                                     includeAllOutcomes = TRUE,
                                     verbosity = "DEBUG")
```

13.7.4 模型开发

在算法的设置功能中，用户可以为每个超参数指定一组合适的值。在训练集中使用交叉验证，通过所谓的网格搜索法，可以遍历所有可能的超参数组合 。如果用户没有指定任何值，则使用默认值。

例如，如果我们对梯度增强学习器使用以下设置：ntrees=c(100, 200）和 maxDepth=4，网格搜索法将为梯度增强机器算法提供两组参数，ntrees=100 和 maxDepth=4 加上其他超参数的默认设置值，ntrees=200 和 maxDepth=4 加上其他超参数的默认设置值。最终模型将会选择交叉验证表现最佳的一组超参数。对于我们的问题，我们选择构建一个具有多个超参数值的 Gradient Boosting Machine (GBM)：

```
gbmModel <- setGradientBoostingMachine(ntrees = 5000,
                                       maxDepth = c(4,7,10),
                                       learnRate = c(0.001,0.01,0.1,0.9))
```

Runplp 函数使用 population、plpData 和 model 设置来训练和评估模型。我们可以使用 testSplit

```
gbmResults <- runPlp(population =population,
                     plpData = plpData,
                     modelSettings = gbmModel,
                     testSplit = 'person',
                     testFraction =0.25,
                     nfold = 2,
                     splitSeed = 1234)
```

（person/time）和 testFraction 参数将数据按 75%-25%拆分，并运行患者水平预测管道模型：

该程序包将在内部使用 R xgboost 包，使用 75%的数据拟合一个 Gradient Boosting Machine 模型，并将在剩余的 25%数据上评估该模型。返回的结果数据包含有关模型及其性能等信息。在 runPlp 函数中，有几个参数用于保存 plpData、plpResults、plpPlots、evaluation 等对象，这些参数默认设置为 TRUE。

我们可以使用以下函数保存模型：

```
savePlpModel(gbmResults$model, dirPath ="model")
```

我们可以使用以下函数加载模型：

```
plpModel <- loadPlpModel("model")
```

还可以使用以下函数保存当前使用的完整的结果结构：

```
savePlpResult(gbmResults, location ="gbmResults")
```

要加载完整的结果结构，请使用以下函数：

```
gbmResults <- loadPlpResult("gbmResults")
```

13.7.5 内部验证

执行研究后，runPlp 函数就会返回训练后的模型以及在训练/测试集上该模型的评估结果。可以通过运行：viewPlp(runPlp=gbmResults)以交互方式查看结果。这将打开一个 Shiny 应用程序，在这个应用中，我们可以查看由框架创建的所有性能评估值，包括交互式图表（请参见 Shiny 应用程序一节中的图 13.16）。

要生成所有评估图表并将其保存到文件夹中，请运行以下代码：

我们在第 13.4.2 节对评估图表进行了更详细的描述。

```
plotPlp(gbmResults, "plots")
```

13.7.6 外部验证

我们建议始终执行外部验证，即在尽可能多的新数据集上应用、评估模型性能。在这里，我们假设第二个数据库上数据提取已经完成并存储在 newData 文件夹中。我们从“模型”文件夹中加载先前拟合的模型：

```
#Load the trained model
plpModel <- loadPlpModel("model")
#Load the new plpData and create the population
plpData <- loadPlpData("newData")
```

```
population <- createStudyPopulation(plpData = plpData,
                                     outcomeId = 2,
                                     washoutPeriod = 364,
                                     firstExposureOnly = FALSE,
                                     removeSubjectsWithPriorOutcome = TRUE,
                                     priorOutcomeLookback = 9999,
                                     riskWindowStart = 1,
                                     riskWindowEnd = 365,
                                     addExposureDaysToStart = FALSE,
                                     addExposureDaysToEnd = FALSE,
                                     minTimeAtRisk = 364,
                                     requireTimeAtRisk = TRUE,
                                     includeAllOutcomes = TRUE)

# apply the trained model on the new data
validationResults <- applyModel(population, plpData, plpModel)
```

为了简化操作，我们还提供了 externalValidatePlp 函数来执行外部验证，同时提取所需的数据。假设我们运行了 result<-runPlp(...)，那么我们可以从中提取模型所需的数据，并在新的数据集上对其进行评估。假设验证队列在表 mainschema.dob.courrent 中，ID 为 1 和 2，CDM 数据在架构 schema.dob 中：

```
valResult <- externalValidatePlp(plpResult = result,
                                 connectionDetails = connectionDetails,
                                 validationSchemaTarget = 'mainschema.dob',
                                 validationSchemaOutcome = 'mainschema.dob',
                                 validationSchemaCdm = 'cdmschema.dbo',
                                 databaseNames = 'new database',
                                 validationTableTarget = 'cohort',
                                 validationTableOutcome = 'cohort',
                                 validationIdTarget = 1,
                                 validationIdOutcome = 2)
```

如果要在多个数据库上验证模型，则可以运行：

```
valResults <- externalValidatePlp(
    plpResult = result,
    connectionDetails = connectionDetails,
    validationSchemaTarget = list('mainschema.dob',
                                  'difschema.dob',
                                  'anotherschema.dob'),
    validationSchemaOutcome = list('mainschema.dob',
                                   'difschema.dob',
                                   'anotherschema.dob'),
    validationSchemaCdm = list('cdms1chema.dbo',
                               'cdm2schema.dbo',
                               'cdm3schema.dbo'),
```

```
    databaseNames = list('new database 1',
                         'new database 2',
                        'new database 3'),
  validationTableTarget = list('cohort1',
                               'cohort2',
                               'cohort3'),
    validationTableOutcome = list('cohort1',
                               'cohort2',
                               'cohort3'),
    validationIdTarget = list(1,3,5),
    validationIdOutcome = list(2,4,6)
)

    validationTableTarget = list('cohort1',
                                 'cohort2',
                                 'cohort3'),
    validationTableOutcome = list('cohort1',
                                  'cohort2',
                                  'cohort3'),
    validationIdTarget = list(1,3,5),
    validationIdOutcome = list(2,4,6)
)
```

13.8 结果外推

13.8.1 模型性能

探索预测模型的性能最简单的做法是使用 viewPlp 函数。这需要一个结果对象作为输入。如果在 R 中开发模型，我们可以使用 runPLp 的结果作为输入。如果使用 ATLAS 生成的研究包，则需要加载其中一个模型（在本例中，我们将加载 Analysis_1）：

```
plpResult <- loadPlpResult(file.path(outputFolder,'Analysis_1', 'plpResult'))
```

这里的“Analysis_1”对应于我们前面指定的分析。

然后，我们可以运行以下命令来启动 Shiny 应用程序。

```
viewPlp(plpResult)
```

Shiny 应用程序打开后会显示该模型在测试和训练集上的性能指标汇总（见图 13.16）。结果表明，训练集的 AUC 为 0.78，而测试集的 AUC 降至 0.74。训练集的 AUC 显示更高的准确率。总的来说，该模型似乎能够在新的 ACE 抑制剂使用者中区分那些将会产生结果的人，但由于训练集的性能优于测试集，该模型有点过拟合。ROC 如图 13.17 所示。

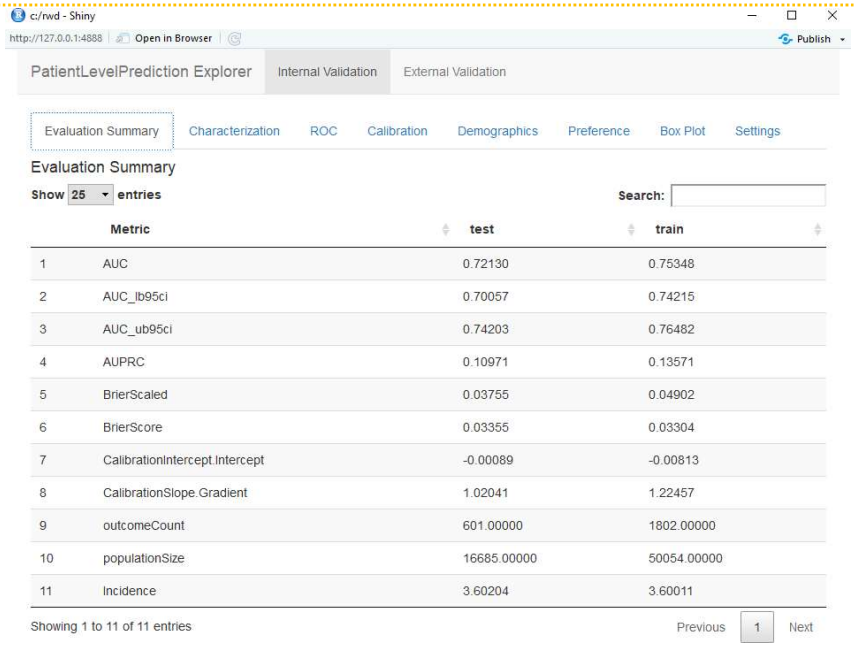


图 13.16:Shiny 应用程序中汇总的评估统计数据。

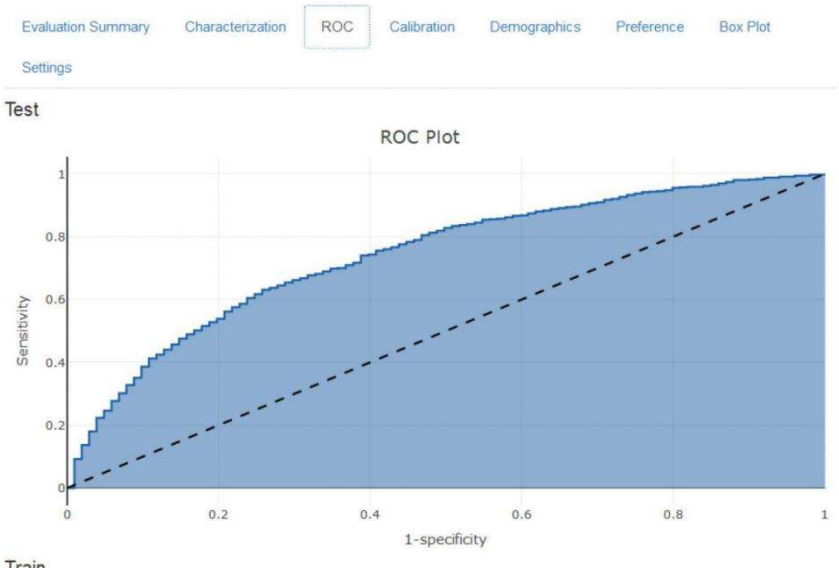


图 13.17: ROC 曲线

图 13.18 中的校准图显示，观察到的风险值总体上与预测的风险值相似，因为点都在对角线周围。然而，图 13.19 中的人口统计学校准图显示，对于年轻患者该模型没有得到很好的校准，因为对于 40 岁以下的患者，曲线（预测风险）与红线（观察风险）不一致。这可能意味着我们需要从目标人群中剔除 40 岁以下的人群（因为观察到的年轻患者的风险几乎为零）。

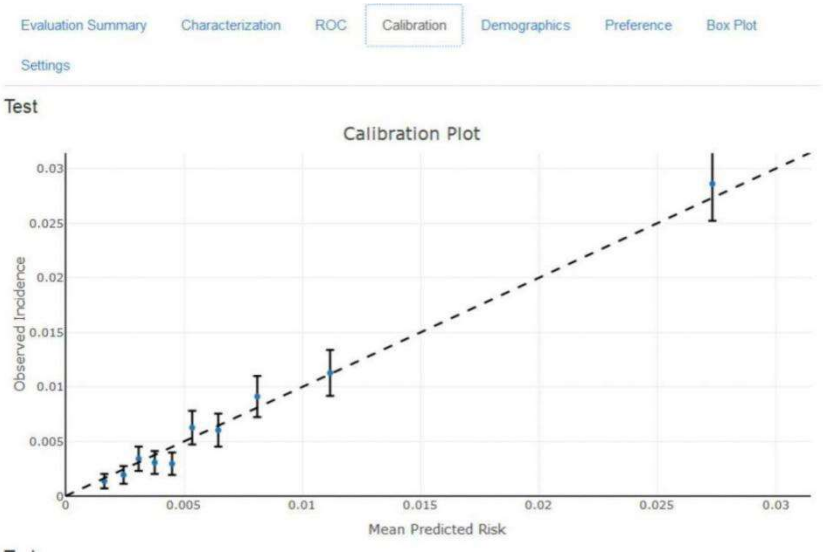


图 13.18：模型校准

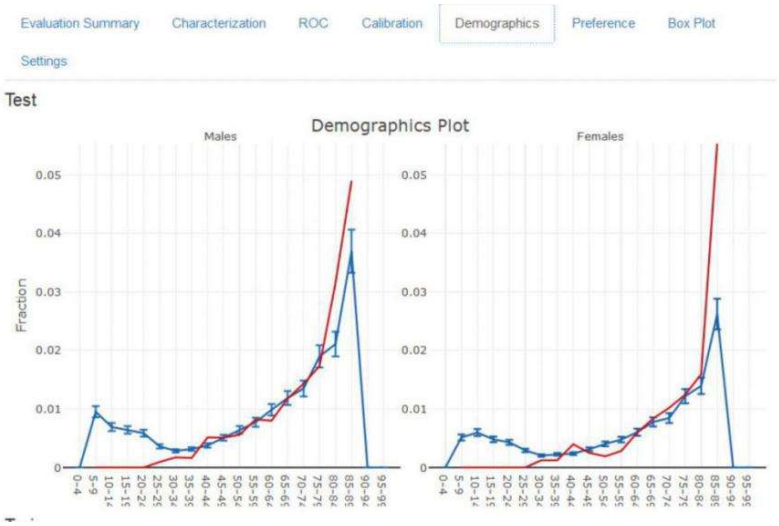


图 13.19：模型的人口统计学校准

最后，损耗图显示了基于纳入/排除标准的标记数据中患者的减少情况（见图 13.20）。该图显示，由于他们没有在完整的风险暴露时间（1 年的随访期）中被持续观察到，我们失去了很大一部分目标人群。有趣的是，有风险发生的人群中缺少完整风险暴露时间观察的人并不多。

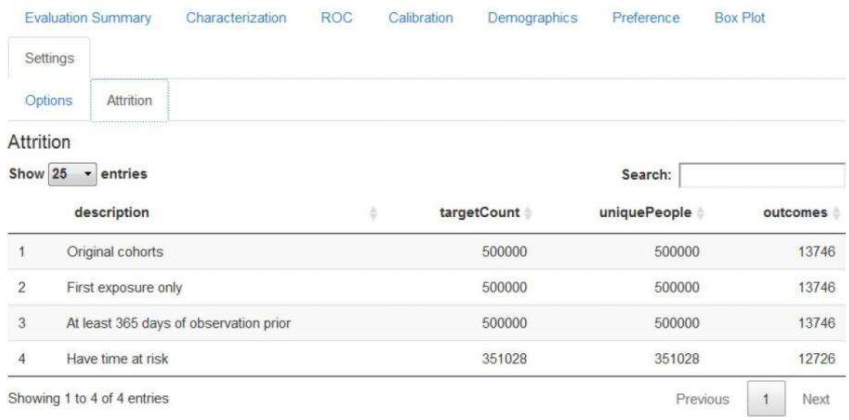


图 13.20：预测问题的损耗图

13.8.2 比较模型

ATLAS 生成的研究包可以针对不同的预测问题生成并评估许多不同的预测模型。因此，专门针对研究包生成的输出，开发了一个额外的 Shiny 应用程序以查看多个模型。要启动此应用程序，请运行 viewMultiplePlp(outputFolder)，其中 outputFolder 是运行 execute 命令时指定的包含分析结果的路径（例如，包含在名为“Analysis_1”的子文件夹中）。

查看模型汇总和设置

交互式 shiny 应用程序将从汇总页面开始，如图 13.21 所示。

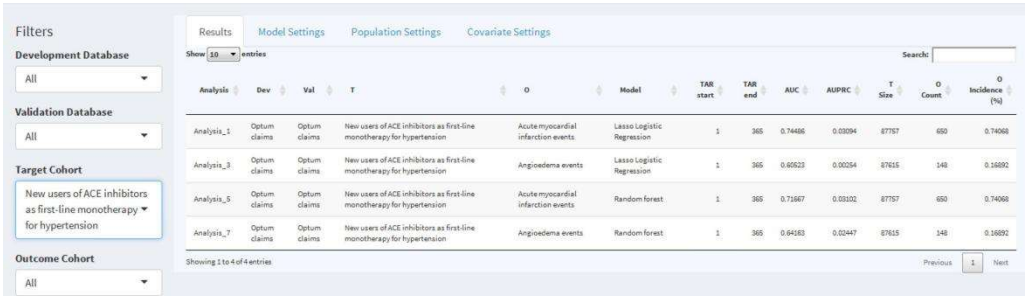


图 13.21：Shiny 汇总页面包含每个训练过的模型的主要性能指标

此汇总页表格包含：

•有关模型的基本信息（例如，数据库信息、分类器类型、风险暴露时间设置、目标群体和结果名称）

•提供目标群体计数和结果发生率

•评估指标： AUC，AUPRC

表格左侧是筛选选项，我们可以在其中指定要关注的开发/验证数据库、模型类型、关注的风险暴露时间设置和/或关注的队列。例如，要选择与目标人群“New users of ACE inhibitors as first line mono-therapy for hypertension”一致的模型，请在 *Target Cohort* 选项中选择该模型。

若要浏览模型，请单击相应的行，选定的行将会高亮显示。选中一行后，我们现在可以通过单击

“Model Settings”选项卡来查看开发模型时使用的模型设置：

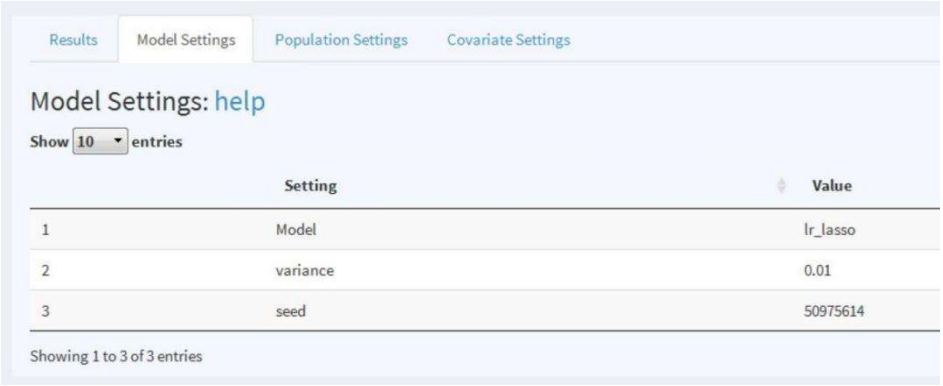
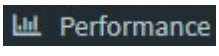


图 13.22：查看开发模型时使用的模型设置。

若要浏览模型，请单击相应的行，选定的行将会高亮显示。选中一行后，我们现在可以通过单击“Model Settings”选项卡来查看开发模型时使用的模型设置：

同样，我们也可以在其他选项卡中查看用于生成模型的群体和协变量设置。

浏览模型性能

选择一行模型后，我们还可以查看其模型性能。单击 

打开阈值性能汇总，如图 13.23 所示：

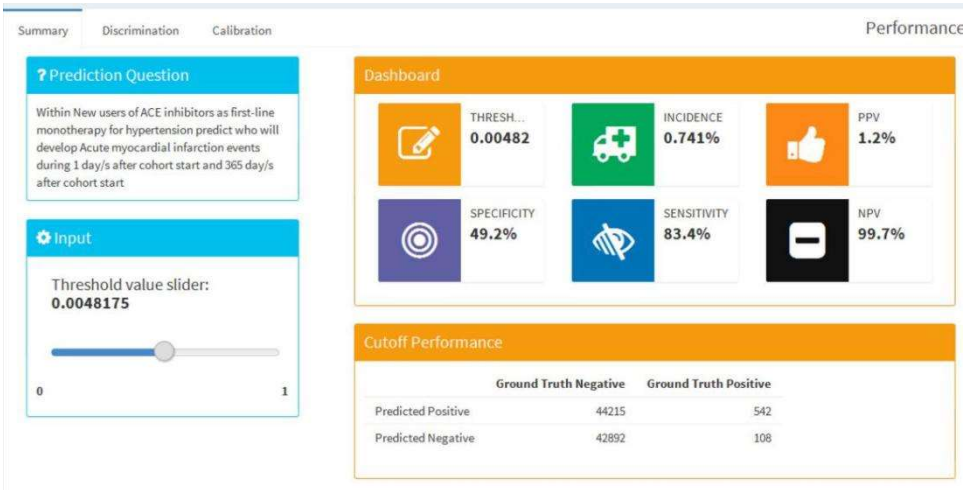


图 13.23：在设定的阈值下的性能指标汇总。

此汇总视图以标准格式显示选定的预测问题、阈值和仪表板，其中包含基于阈值的关键指标，如阳性预测值（PPV）、阴性预测值（NPV）、敏感性和特异性（请参阅第 13.4.2 节）。在图 13.23 中，我们看到在阈值为 0.00482 的情况下，敏感性为 83.4%（在次年发生风险的患者中 83.4%的患者的预后风险大于或等于 0.00482），PPV 为 1.2%（预后风险大于或等于 0.00482 的患者中 1.2%的患者在下一年发生了风险）。由于一年内风险的发生率为 0.741%，识别预后风险大于或等于 0.00482 的患者可以发现亚组患者的风险几乎是普通人群平均风险的两倍。我们可以使用滑块调整阈值，以查看模型在其他

阈值下的性能表现。

要查看模型的总体差异，请单击“差异”选项卡以查看 ROC 图、准确率-召回率图和分布图。曲线图上的线对应于选定的阈值点。图 13.24 显示了 ROC 图和准确率-召回率图。ROC 曲线图显示该模型能够区分年内将会发生风险和不会发生风险的概率。然而，当我们看到准确率-召回率图时，其性能看起来并不可观，因为风险的低发生率可能意味着有很高的假阳性率。

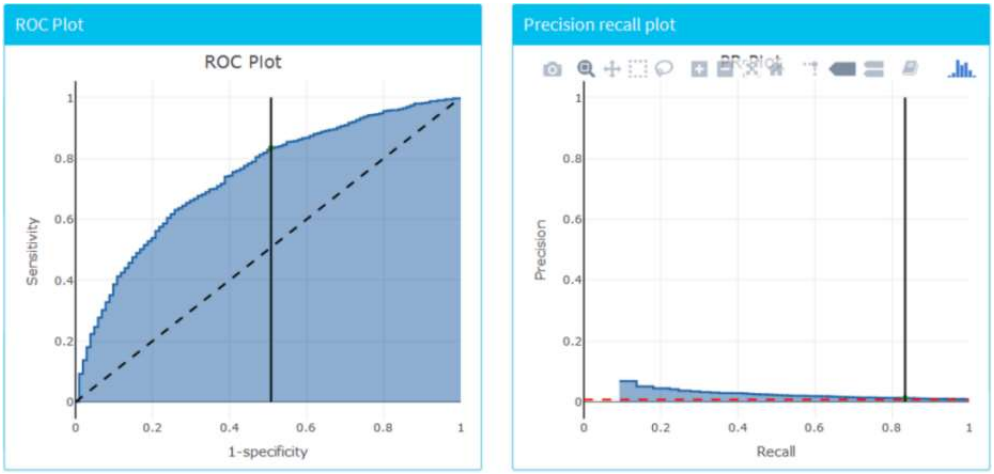


图 13.24：用于查看模型总体辨别能力的 ROC 图和准确率-召回率图。

图 13.25 显示了预测和偏好得分的分布情况。

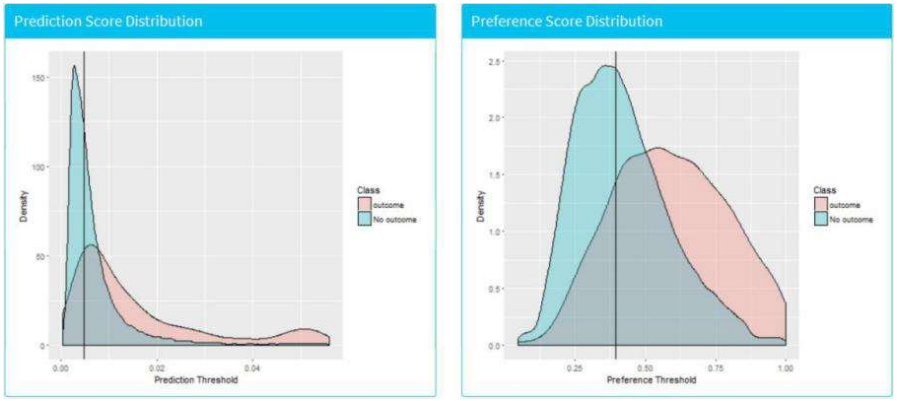


图 13.25：有或无结果群体的预测风险分布。重叠区域越多，则辨别力越差。

最后，我们还可以通过单击“校准”选项卡来查看模型的校准情况。这将显示校准图和人口统计学校准图，如图 13.26 所示。

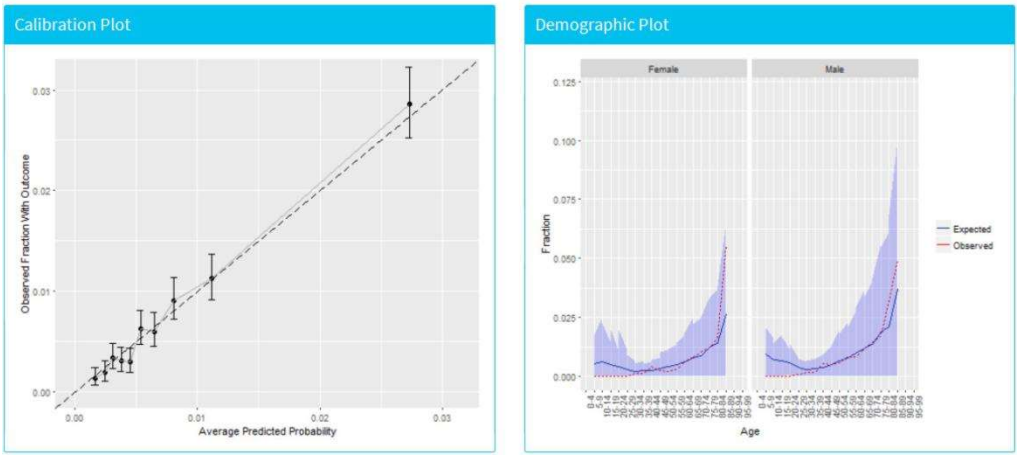
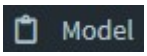


图 13.26：风险等级校准图和人口统计学校准图

我们可以看到平均预测风险似乎与经过一年实践的观察分数相匹配，因此该模型得到了很好的校准。有趣的是，人口统计学校准图显示，年轻患者的预期风险高于观察到的风险水平，因此我们对年轻年龄组的预测风险准确率偏高。相反，对于 80 岁以上的患者，该模型预测的风险低于观察到的风险水平。这种情况可能促使我们为年轻或年长患者开发独立的模型。

查看模型

要检查最终模型，请从左侧菜单中选择 

选项。这将打开一个界面，其中包含模型中每个变量的图形，如图 13.27 所示，以及一个汇总所有候选协变量的表格，如图 13.28 所示。变量图分为二元变量和连续型变量。X 轴是非结果队列的患者的患病率/平均值，Y 轴是有结果的患者的患病率/平均值。我们可以看到，对于有结果的患者，变量的任意点更多地落在对角线的上方，较少会落在对角线的下方。

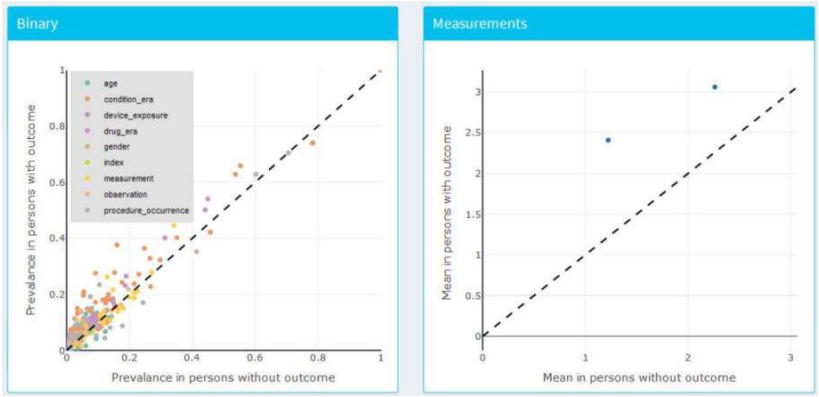


图 13.27：模型汇总图。每种点对应一个模型中包含的变量

图 13.28 中的表格显示了所有候选协变量的名称、值（如果使用通用线性模型，则为系数，否则为变量重要性）、有结果平均值（结果队列的平均值）和无结果平均值（非结果队列的平均值）。

Model Table

Download Model

Show 10 entries Search:

	Covariate Name	Value	Outcome Mean	Non-outcome Mean
1	age group: 00-04	0	0.0004	0.0001
2	age group: 05-09	0	0	0.0003
3	index month: 1	0	0.1307	0.1096
4	observation during day -365 through 0 days relative to index: Domain	0	0.1188	0.0514
5	Charlson index - Romano adaptation	0	2.4783	1.3817
6	Diabetes Comorbidity Severity Index (DCSI)	0.1478	2.4056	1.2207
7	CHADS2VASc	0.9279	3.0573	2.2576
8	visit_occurrence concept count during day -365 through 0 concept_count relative to index	0	19.5263	13.8837
9	age group: 10-14	0	0	0.001
10	index month: 2	0	0.0934	0.0909

Showing 1 to 10 of 67,897 entries Previous 1 2 3 4 5 … 6790 Next

图 13.28：模型详细表格。

预测模型不是因果模型，不应将预测变量误认为原因。无法保证修改图 13.28 中的任何变量都会对结果的风险产生影响。

13.9 其他患者水平预测功能

13.9.1 期刊论文生成

我们添加了自动生成 Word 文档的功能，可以用作期刊论文的初稿。它包含了许多已生成的研究详细信息和结果。如果我们进行了外部验证，这些结果也可以添加进来。或者我们可以添加一个“Table 1”，其中包含目标群体的许多协变量的数据。我们可以通过运行以下函数来创建期刊论文草稿：

```
createPlpJournalDocument(plpResult = <your plp results>,
                         plpValidation = <your validation results>,
                         plpData = <your plp data>,
                         targetName = "<target population>",
                         outcomeName = "<outcome>",
                         table1 = F,
                         connectionDetails =NULL,
                         includeTrain = FALSE,
                         includeTest = TRUE,
                         includePredictionPicture = TRUE,
                         includeAttritionPlot = TRUE,
                         outputLocation = "<your location>")
```

有关更多详细信息，请参见该函数的帮助页面。

13.10 总结

患者水平预测旨在开发一个利用历史数据预测未来事件的模型。

选择用于模型开发的最优算法是一个经验问题，即它应该由问题和手头的数据决定。

在利用存储于 OMOP-CDM 中的数据进行预测模型开发和验证时，PatientLevelPrediction 包实现了最佳实践。

可以通过交互式面板实现模型及其性能评估的展现

OHDSI 的预测框架能够实现预测模型的大规模外部验证，这是临床采用的先决条件。

13.11 练习

先决条件

对于这些练习，假设 R、R-Studio 和 Java 已经按照第 8.4.5 节的描述安装。还需要 SqlRender、DatabaseConnector、Eunomia 和 PatientLevelPrediction 包，这些包可以使用以下方法安装：

```
install.packages(c("SqlRender", "DatabaseConnector", "devtools"))
devtools::install_github("ohdsi/Eunomia", ref = "v1.0.0")
devtools::install_github("ohdsi/PatientLevelPrediction")
```

Eunomia 包在 CDM 中提供了一个模拟的数据集，该数据集将在本地 R 会话中运行。该数据集的连接详细信息可通过以下方式获得：

CDM 数据库架构是“main”。这些练习会使用到几个队列。Eunomia 包中的 createColorts 函数将在表 Cohort 中创建这些队列：

问题定义

在首次使用 NSAIDs 的患者中，预测将在明年出现胃肠道出血的群体。

NSAID 新用户队列的 COHORT_DEFINITION_ID 为 4。胃肠道出血队列的 COHORT_DEFINITION_ID 为 3。

练习 13.1. 使用 R 包 PatientLevelPrediction，定义要用于预测的协变量，并从 CDM 中提取 PLP 数据。创建 PLP 数据的摘要。

```
Eunomia::createCohorts(connectionDetails)
```

练习 13.2. 重新浏览定义最终目标群体所需的设计选择，并使用 createStudyPopulation 函数指定这些选择。你的选择会对目标群体的最终规模产生什么影响?

练习 13.3. 使用 LASSO 建立一个预测模型，并使用 Shiny 应用程序评估其性能。你的模型性能表现如何？建议的答案可在附录 E.9 中找到。

参考文献

26. Byrd, J. B., A. Adam, and N. J. Brown. 2006. “Angiotensin-converting enzyme inhibitor-associated angioedema.” Immunol Allergy Clin North Am 26 (4): 725–37.
27. Cicardi, M., L. C. Zingale, L. Bergamaschini, and A. Agostoni. 2004. “Angioedema associated with angiotensin-converting enzyme inhibitor use: outcome after switching to a different treatment.” Arch. Intern. Med. 164 (8): 910–13.
28. Norman, J. L., W. L. Holmes, W. A. Bell, and S. W. Finks. 2013. “Life-threatening ACE inhibitor-induced angioedema after eleven years on lisinopril.” J Pharm Pract 26 (4): 382–88.
29. O’Mara, N. B., and E. M. O’Mara. 1996. “Delayed onset of angioedema with angiotensin-converting enzyme inhibitors: case report and review of the literature.” Pharmacotherapy 16 (4): 675–79.
30. Reps, J. M., M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek. 2018. “Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data.” Journal of the American Medical Informatics Association 25 (8): 969–75. https://doi.org/10.1093/jamia/ocy032.
31. Thompson, T., and M. A. Frable. 1993. “Drug-induced, life-threatening angioedema revisited.” Laryngoscope 103 (1 Pt 1): 10–12.https://www.equator-network.org/reporting-guidelines/tripod-statement/

第十四章 证据质量

章节负责人：Patrick Ryan 和 Jon Duke

14.1 可信证据的属性

任何旅程开始前对理想终点的展望都可以是很有帮助的。为了支撑数据和证据之间的关联，我们强调让证据质量可信的属性。

可信的证据应该是可重复的（repeatable）。这意味着研究人员针对任何特定问题，使用相同方法对相同数据进行分析时，应该得到完全一致的结果。这个最基本要求隐含的概念是，证据是使用特定输入并执行一系列规定程序的结局，且在这个过程开始后不需要任何人工干预。更理想的情况是，可信的证据应该是可重现的(reproducible)。也就是说，不同的研究人员应该能够完成相同的任务，基于指定的数据库，执行指定的分析，得到与第一个研究人员相同的结果。

所需	研究问	研究	数据	分析	结论
可重	一致	一致	一致	一致	一致
可重	一致	不同	一致	一致	一致
可复	一致	相同	相似	一致	相似
可外	一致	相同	不同	一致	相似
稳健	一致	相同	相同	不同	相似
被校	相似	一致	一致	一致	统计

图 14.1:可信的证据所需要具备的属性

可重现性要求产生证据的过程是完全指定的，这个过程可以以人类可读和计算机可执行的形式存在，目的是避免研究者对研究过程进行人为干预。实现可重复性和可重现性的最有效解决方法是使用严格定义了输入和输出的标准化分析路径，并将该标准化分析路径应用于指定版本的数据库。

如果对相同的问题使用相同的分析方法处理分析类似的数据后可以产生相似的结论，那么这种证据被认为是**可复制的（replicable）**，从而提高证据的可信性。例如，在一个大型私人保险公司的医保数据库进行分析所产生的证据，如果在其他保险公司的医保数据上得以验证，那么这个证据的可信性就会得到加强。在群体水平效果评估研究中，可复制性这一属性符合奥斯汀·布拉德福德·希尔爵士关于一致性的因果论点。“是否被不同的人在不同的地点、环境和时间被反复观测到?.....对于所发现的风险究竟是真实的还是纯属偶然，有时可能只能通过环境重现和重复观察来回答（Hill, 1965)。”在患者水平的预测任务中，可复制性则在于将基于一个数据库训练得到的模型用于别的数据库时应该得到相似的准确性和并能根据不同的数据库进行校正。如果对不同的数据库执行相同的分析能得到相对一致的结论，在这种情况下，我们可以进一步确信证据是**可外推的（generalizable)**。OHDSI 研究网络的一个核心价值在于能够提供不同的人群、地理位置和数据收集过程所带来的多样性。Madigan 等人(2013b)的研究表明，效应估计对数据的选择是敏感的。每个数据源都有其固有的局限性和特定偏倚，从而限制了单一发现的可信度。因此，如果在异质的数据集中观察到相似模式，这样的发现就是强有力的，因为重复的发现大大降低了特定数据源偏倚造成该发现的可能性。如果在美国、欧洲和亚洲的多个医保和电子病历数据库中对人群水平效应估计的结果显示一致，那么它们应该被视为证明医学干预有效的有力证据，因此可以在更大范围内影响医疗决策。

可信的证据应该是**稳健的(robust)**，表示研究结论不应该对分析过程中做出的主观选择过于敏感。如果某项研究可以使用其他合理的统计方法，那么如果不同方法产生相似结论，则为该证据提供进一步保证；反之，如果产生不同结论，则为该证据的可信度产生警示（Madigan, Ryan, and Schuemie 2013)。对于群体效应估计，敏感性分析可以在两个层级进行，第一层级是在研究设计阶段采取不同的研究方案，比如采用比较队列或采用自身对照病例系列设计。第二层级是在研究设计内嵌的数据分析层面，比如在比较队列框架设计内采用倾向评分匹配、分层或权重作为调整混杂的策略。

最后，也可能是最重要的一点，证据应当是能够**被校验的（calibrated)**。如果一个证据生成系统的性能不能被验证，那么这个系统是无法用于对未知问题产生答案的。一个闭合的系统应该具有已知的运行特性，这些特性应当能被测量并作为背景用于解释该系统产生的任何结果。统计假象应具有明确定义的属性。例如，95%置信区间具有 95%的覆盖率，或者预测概率为 10%的队列应该只在10%的人群中观察到事件。观察性研究应该总是伴随着研究诊断，即对设计、方法和数据的假设进行检验。这些诊断应着重于评估影响研究有效性的主要因素：选择偏差、混杂和测量误差。阴性对照已被证明是识别和减少观察性研究中系统性错误的有力工具（Schuemie 等， 2016; Schuemie, Hripcsak, 等， 2018; Schuemie, Ryan, 等， 2018)。

14.2 理解证据质量

但是，我们如何知道一项研究的结论是否足够可信呢？它们能被信任用于临床吗？在监管决策方面呢？它们可以作为未来研究的基础吗？每当一项新的研究发表或传播时，读者都必须考虑这些问题，无论该研究是随机对照试验、观察性研究还是其他类型的分析。

在观察性研究和“真实世界数据”的应用中经常被关注的问题之一是数据质量(Botsis 等，2010；

Hersh 等，2013；Sherman 等，2016)。通常来说，用于观察性研究的数据最初并不是为了当下的研究目的而收集的，因此可能存在数据获取的不完整或不准确以及固有偏倚。这些问题已经引起了越来越多关于如何度量、描述和改进数据质量的研究（Kahn 等，2012；Liaw 等，2013；Weiskopf 和 Weng，2013)。OHDSI 社区是这类研究的强烈倡导者，社区成员牵头组织并参与了许多关于 OMOP 通用数据模型和 OHDSI 网络中数据质量的研究（Huser 等，2016；Kahn 等，2015；Callahan 等，2017；Yoon 等，2016)。

从过去十年中该领域的发现可以看到，数据质量明显不是，也永远不会是完美的。这一观点在医学信息学领域的先驱克莱姆·麦克唐纳博士的论述中得到了很好的体现：

精准度的遗失开始于数据从医生大脑转移到医疗记录的时刻。

因此，作为一个群体，我们必须提出这样一个问题：面对不完美的数据，我们如何才能获得可信的证据？

答案就在于对“证据质量”的整体观：检验从数据到证据的整个过程，明确证据产生过程的每个组成部分，衡量每个组成部分的质量，并且针对在此过程中每一步学习到的东西进行坦诚的沟通。证据质量不仅应该考虑观察性数据的质量，也要考虑在观察性分析中使用的研究方法、软件和临床定义的有效性。

在接下来的章节中，我们将探讨表 14.1 中列出的证据质量的四个组成部分。

表 14.1：证据质量的四个组成部分

证据质量的组成部分	测量内容
数据质量	数据是否以符合协定的结构和约定的方式，以合理的值进行完整采集？
临床有效性	进行的分析在多大程度上符合临床目的？
软件有效性	我们能相信转换和分析数据的过程完成了其应当做的事情吗？
研究方法有效性	考虑到数据的优缺点，这个研究方法适合这个研究问题吗？

14.3 沟通证据质量

衡量证据质量的一个重要方面是能够展示从数据到证据的过程中产生的不确定性。OHDSI 解决证据质量的总体目标是使医疗决策者对 OHDSI 生成的证据树立信心。尽管在很多方面毋庸置疑还不完善，但是这些证据的缺点和优点一直被持续监测，并对此类信息以严谨而开放的方式进行沟通。

14.4 总结

生成的证据应当是可重复、可重现、可复制、可外推、稳健和被校验的。

当回答证据是否可信这一问题时，证据质量应考虑包括数据质量的各个方面：

-数据质量

-临床有效性

-软件有效性

-研究方法有效性

当沟通证据时，应当展示因为证据质量不完美而带来的不确定性.

参考文献

32. Botsis, Taxiarchis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. 2010. “Secondary Use of Ehr: Data Quality Issues and Informatics Opportunities.” *Summit on Translational Bioinformatics* 2010: 1.
33. Callahan, Tiffany J, Alan E Bauck, David Bertoch, Jeff Brown, Ritu Khare, Patrick B Ryan, Jenny Staab, Meredith N Zozus, and Michael G Kahn. 2017. “A Comparison of Data Quality Assessment Checks in Six Data Sharing Networks.” *eGEMs* 5 (1).
34. Hersh, William R, Mark G Weiner, Peter J Embi, Judith R Logan, Philip RO Payne, Elmer V Bernstam, Harold P Lehmann, et al. 2013. “Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research.” *Medical Care* 51 (8 0 3): S30.
35. Hill, A. B. 1965. “THE ENVIRONMENT AND DISEASE: ASSOCIATION OR CAUSATION?” *Proc. R. Soc. Med.* 58 (May): 295–300.
36. Huser, Vojtech, Frank J. DeFalco, Martijn Schuemie, Patrick B. Ryan, Ning Shang, Mark Velez, Rae Woong Park, et al. 2016. “Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets.” *EGEMS (Washington, DC)* 4 (1): 1239. https://doi.org/10.13063/2327-9214.1239.
37. Kahn, Michael G., Jeffrey S. Brown, Alein T. Chun, Bruce N. Davidson, Daniella Meeker, P. B. Ryan, Lisa M. Schilling, Nicole G. Weiskopf, Andrew E. Williams, and Meredith Nahm Zozus. 2015. “Transparent Reporting of Data Quality in Distributed Data Networks.” *EGEMS (Washington, DC)* 3 (1): 1052. https://doi.org/10.13063/2327-9214.1052.
38. Kahn, Michael G, Marsha A Raebel, Jason M Glanz, Karen Riedlinger, and John F Steiner. 2012. “A Pragmatic Framework for Single-Site and Multisite Data Quality Assessment in Electronic Health Record-Based Clinical Research.” *Medical Care* 50.
39. Liaw, Siaw-Teng, Alireza Rahimi, Pradeep Ray, Jane Taggart, Sarah Dennis, Simon de Lusignan, B Jalaludin, AET Yeo, and Amir Talaei-Khoei. 2013. “Towards an Ontology for Data Quality in Integrated Chronic Disease Management: A Realist Review of the Literature.” *International Journal of Medical Informatics* 82 (1): 10–24.
40. Madigan, D., P. B. Ryan, and M. Schuemie. 2013. “Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies.” *Ther Adv Drug Saf* 4 (2): 53–62.
41. Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. 2013. “Evaluating the impact of database heterogeneity on observational study results.” *Am. J. Epidemiol.* 178 (4): 645–51.
42. Schuemie, M. J., G. Hripcsak, P. B. Ryan, D. Madigan, and M. A. Suchard. 2016. “Robust empirical calibration of p-values using observational data.” *Stat Med* 35 (22): 3883–8.

43. Schuemie, M. 2018. “Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data.” *Proc. Natl. Acad. Sci. U.S.A.* 115 (11): 2571–7.
44. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. “Improving reproducibility by using high-throughput observational studies with empirical calibration.” *Philos Trans A Math Phys Eng Sci* 376 (2128).
45. Sherman, Rachel E, Steven A Anderson, Gerald J Dal Pan, Gerry W Gray, Thomas Gross, Nina L Hunter, Lisa LaVange, et al. 2016. “Real-World Evidence—What Is It and What Can It Tell Us.” *N Engl J Med* 375 (23): 2293–7.
46. Weiskopf, Nicole Gray, and Chunhua Weng. 2013. “Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research.” *Journal of the American Medical Informatics Association: JAMIA* 20 (1): 144–51. https://doi.org/10.1136/amiajnl-2011-000681.
47. Yoon, D., E. K. Ahn, M. Y. Park, S. Y. Cho, P. Ryan, M. J. Schuemie, D. Shin, H. Park, and R. W. Park. 2016. “Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research.” *Healthc Inform Res* 22 (1): 54–58.

第十五章 数据质量

章节负责人：Martijn Schuemie，Vojtech Huser，Clair Blacketer

大部分观察性医疗健康数据并不是以研究为目的而进行收集的。例如，电子健康档案（electronic health records, EHR）主要记录关键信息以便为患者提供诊疗服务，医保报销数据主要用于提供报销依据。很多学者质疑将此类数据用于临床研究是否合适，van der Lei (1991) 甚至指出“数据应仅用于其收集的目的”，担心数据收集非源于研究目的，不能确保它的质量。如果收集的数据质量很差（垃圾数据输入），那么使用该数据得出的研究结果的质量也必定很差（垃圾结果输出）。因此，医疗健康领域观察性研究中的一个重要问题是评估数据质量，目的在于回答下面这个问题：

所收集数据的质量能否满足我们的研究目的?

参考既往文献(Roebuck,2012)，可将数据质量（data quality）定义为:

数据能够用于特定用途，且具有完整性、有效性、一致性、及时性和准确性。

请注意，数据不大可能是完美的，但保障其质量对于某个特定的研究目的而言可能已经足够了。

我们虽然无法直接观察到数据质量，但已经开发了评估数据质量的方法。目前有两种不同的数据质量评估方法(Weiskopf and Weng, 2013)：一般评估法和特定研究背景评估法。

在本章中，首先汇总数据质量问题的可能来源，然后讨论一般评估法和特定研究背景评估法的理论，最后逐步说明如何使用 OHDSI 工具评估数据质量。

15.1 数据质量问题的来源

影响数据质量的过程有很多。正如第 14 章中提到的，当医生开始记录她/他的想法时，就已经开始出现数据质量降低的问题了。Dasu and Johnson (2003) 按照下面类别，建议将数据质量管理融入到数据生命周期的每个环节中，并将其称为数据质量的连续体：

1. **数据采集和整合：**数据质量问题可能源于手动输入错误、偏倚 (如报销时虚报医药费)、电子健康档案中表格的错误连接以及使用默认值替换缺失值等
2. **数据存储和知识共享：**潜在的数据质量问题源于缺乏规范的数据模型归档和元数据。
3. **数据分析：**数据质量问题包括不正确的数据转换、不正确的数据解读以及使用不适当的分析方法。
4. **数据发布**：数据质量问题可能源于发布数据供后续使用的过程中。

通常，我们使用的数据是已经完成采集和整合的，因此对于步骤 1 数据质量的改善几乎无能为力。当然我们确实有方法检查此步骤的数据质量，这将在本章的后续章节中进行阐述。

同样，我们通常接收的数据已有特定的形式，因此我们对步骤 2 数据质量的改善所能做的工作也很有限。但在 OHDSI 中，我们会将所有观测数据转换为通用数据模型（Common Data Model，CDM），这个过程我们是有能力把控其质量的。有研究者担心此步骤也会降低数据质量。由于这一过程由我们控

制，因此可以通过建立严格的保护措施来确保数据质量，这部分内容将在本节 15.2.2 进行阐述。已有几项调查显示(Defalco et al., 2013; Makadia and Ryan, 2014; Matcho et al., 2014; Voss et al., 2015a,b; Hripcsak et al., 2018)，将数据正确的转换为 CDM 后，几乎不会出现任何影响数据质量的问题。事实上，通过大型数据管理社区共享的文档化数据模型来处理数据将利于数据以更加明确而清晰的方式存储。

步骤 3（数据分析）也可以由我们把控。在 OHDSI 中，关于数据分析的质量问题，我们一般不建议使用数据质量这一术语，而是使用临床有效性、软件有效性和方法有效性等术语，这将在第 16,17,18 章中详细阐述。

15.2 一般数据质量

我们可以问自己一个问题，我们的数据是否符合观察性研究的一般目的。Kahn 等(2016)将一般数据质量分为三个部分：

1. **一致性：**数据的值是否符合指定的标准和格式？一般，数据类型分为三种：

- **值：**记录的数据是否与指定的格式一致？例如，所有医疗专业都是有效专业吗？
- **关系：**记录的数据是否符合指定的约束关系？例如，DRUG_EXPOSURE 数据库中的 PROVIDER_ID 变量是否在 PROVIDER 表中有相对应的记录？
- **计算：**通过对数据进行计算是否会产生预期的结果？例如，根据身高和体重计算出的 BMI 值，是否与数据库中存在的 BMI 值一致？

2. **完整性：**是否缺失特定变量（例如，在诊室测量的体重是否被记录了？）以及变量是否包含所有记录的值（例如，所有人的性别都已知吗？）

3. **合理性：**数据可信吗？可以从三个角度来评估：

- **唯一性：**例如，每个 PERSON_ID 值在 PERSON 表中仅出现一次吗？
- **非时间性：**数据的值、分布或密度是否与期望一致？例如，数据库计算的糖尿病患病率是否与已知患病率无差别？
- **时间性：**数据值的变化是否符合预期？例如，疫苗接种顺序是否与临床指南所推荐的顺序相一致？

每一部分都可以通过两种方式来评估：

- **核实：**重点在于模型和元数据的数据约束、系统假设以及本地知识，而不依赖外部参照基准。核实的关键功能是，能够利用本地环境中的资源确定期望值和分布。
- **验证：**着眼于数据值是否与外部参照基准一致。外部参照基准的来源之一可能是合并多个数据集的结果。

15.2.1 数据质量检查

Kahn 引入了“数据质量检查（data quality check）”这一术语，有时称为“数据质量规则（data quality rule）”，用于检查数据是否符合指定要求（例如，标记出患者年龄 141 岁的异常值，可能是由于出生年份不正确或死亡事件缺失引起）。通过在软件中创建自动数据质量检查工具可以实现这种数据检查，如大规模纵向证据系统中健康资料的自动化特征描述工具(Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems，ACHILLES) (Huser et al., 2018)。ACHILLES 是一个软件工具，可对符合 CDM 的数据库进行特征描述和可视化分析。因此，它

可用于评估数据库网络中的数据质量(Huser et al., 2016) 。ACHILLES 不仅可以作为独立工具使用，也可以作为“数据源”函数集成到 ATLAS 中使用。

ACHILLES 预先计算 170 多个数据的描述性分析，每个分析过程都有一个分析 ID 编码和该分析的简短描述；例如“ 715：按 DRUG_CONCEPT_ID 分类所得 DAYS_SUPPLY 的分布”和“ 506：按性别分类所得死亡年龄”。这些分析结果将存储在数据库中，并且可以通过 Web 查看器或 ATLAS 获得。

OHDSI 社区创建的另一种评估数据质量的工具是数据质量仪表板（Data Quality Dashboard，DQD）。ACHILLES 进行的描述分析可整体可视化地理解一个 CDM 实例，而 DQD 逐表逐字段地量化 CDM 中与给定规范不符的记录数。总共进行 1,500 多次检查，每次都有序地汇总到 Kahn 框架中。每次检查均会将结果与阈值进行比较，当违规行所占百分比超出该阈值时则被认为是失败（FAIL）。表 15.1 提供了一个检查实例。

表15.1：数据质量仪表板中的数据质量检查

违规行百分比	检查说明	阈值	状态
0.34	是或否表示按照规则provider_id在VISIT_OCCURRENCE是否为预期数据类型。	0.05	失败
0.99	MEASUREMENT表中映射为0的measurement_source_value字段不同来源值的记录数和百分比。	0.30	失败
0.09	DRUG_ERA表中的drug_concept_id字段中的值不符合组成类别的记录数和百分比	0.10	通过
0.02	DRUG_EXPOSURE中的verbatim_end_date字段（日期）发生在DRUG_EXPOSURE表中DRUG_EXPOSURE_START_DATE字段（日期）之前的记录数和百分比。	0.05	通过
0.00	PROCEDURE_OCCURRENCE中的procedure_occurrence_id字段中出现重复值的记录数和百分比。	0.00	通过

DQD 工具检查有多种形式，其中一种是表、字段和概念级别的检查。表级别的检查是 CDM 中的高层检查，例如确认是否所有必需的表都存在；字段级别检查是评估每个表中的每个字段是否符合 CDM 规范，包括确保所有主键都是真正唯一的，并且所有标准概念字段都在适当的域中包含 ID 概念，以及其他内容；概念级别的检查会更深入地检查各个 ID 概念。其中许多措施属于 Kahn 框架的合理性类别，例如确保性别相关的特定概念不会出现性别错误（例如：前列腺癌出现在女性患者中）。

ACHILLES 和 DQD 依托 CDM 数据运行。数据质量问题可能是在将数据转换为 CDM 数据时出现，但也可能是原始数据本就存在问题。如果是转换有误，此类问题通常是可以控制的，但如果数据本身就有问题，

15.2.2 ETL 检查单元

除了高级数据质量检查之外，还应该进行个体级别的数据检查。ETL 过程（即提取-转换-加载）是将数据转换为 CDM 数据的过程，这个过程通常非常复杂，随之而来的是犯错误的风险也在增加，且这些错误很容易被忽视。另外，随着时间的推移，源数据模型可能会更改，或者 CDM 数据可能会更新，因此有必要对 ETL 过程进行修改。然而，ETL 过程比较复杂，如果进行更改很有可能产生意想不到的后果，因此需要重新考虑和审查 ETL 的所有方面。

为了确保 ETL 能够按照预期目标进行，并可以持续进行，强烈建议构造一组检查单元。检查单元是一小段代码，可以自动检查某个内容。第 6 章中介绍的 Rabbit-in-a-Hat 工具可以创建一个检查单元框架，从而简化对此类检查单元代码的编写。该框架是 R 函数的集合，专门为 ETL 的源数据库和目标 CDM 数据版本创建。其中一些功能可用于创建符合源数据模式的伪数据记录，而其他功能可用于指定 CDM 格式数据的期望。下面是一个检查单元示例：

```
source("Framework.R")
declareTest(101, "Person gender mappings")
add_enrollment(member_id = "M000000102", gender_of_member = "male")
add_enrollment(member_id = "M000000103", gender_of_member = "female")
expect_person(PERSON_ID = 102, GENDER_CONCEPT_ID = 8507
expect_person(PERSON_ID = 103, GENDER_CONCEPT_ID = 8532)
```

在上面这个例子中，由 Rabbit-in-a-Hat 生成的框架是基础，并加载其余代码中的函数。然后，我们给出指令开始检查人的性别映射。源模式有一个 ENROLLMENT 表，我们使用 Rabbit-in-a-Hat 创建的 add_enrollment 函数为 MEMBER_ID 和 GENDER_OF_MEMBER 字段创建两个具有不同值的记录。最后，我们指定了以下期望：在 ETL 之后，PERSON 表中应该存有两个记录，且具有不同的期望值。

请注意，ENROLLMENT 表还有许多其他字段，但在此检查中，我们不太关心其他字段的值。不过，如果将这些值留空（例如出生日期）可能会导致 ETL 误删这一记录或引发错误。为了克服这个问题，同时使检查代码易于阅读，add_enrollment 函数将为用户未明确指定的字段设定为默认值（即“White Rabbit”扫描报告中观察到的最普遍的值）。

可以为 ETL 中所有其他逻辑单元创建类似的检查单元，一般会产生数百次检查。定义完检查后，可以使用该框架生成两组 SQL 语句，一组用于创建伪源数据，另一组用于创建 ETL 版本数据：

```
insertSql <- generateInsertSql(databaseSchema = "source_schema")

testSql <- generateTestSql(databaseSchema = "cdm_test_schema")
```

如图 15.1 展示了整个流程。

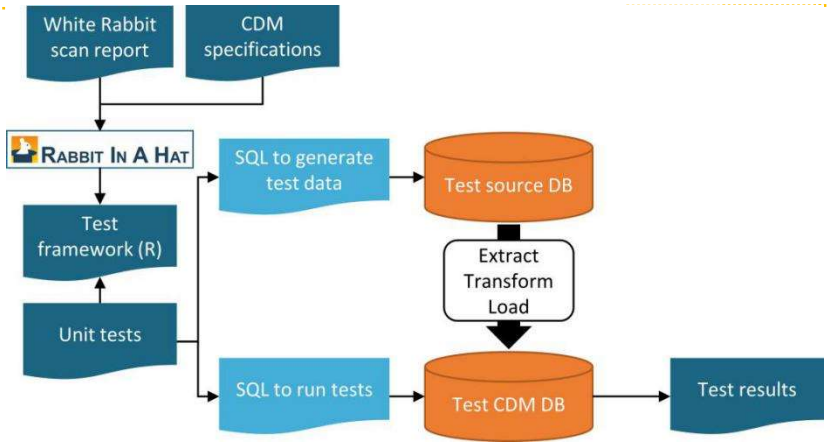


图 15.1：使用 Rabbit-in-a-Hat 检查框架对 ETL 过程（提取-转换-加载）进行的单元检查。

SQL 检查返回的表类似表 15.2，前面定义的两个检查均顺利通过。

表 15.2：ETL 单元检查结果示例

ID	描述	状态
101	人的性别映射	通过
101	人的性别映射	通过

单元检查的强大之处在于，ETL 流程更改后，仍然可以轻松地在任何时间运行它们。

15.3 研究专用检查

到目前为止，本章着重于常规数据质量(DQ)检查，应在将数据用于研究之前执行此类检查，由于这些检查与所研究的问题无关，因此我们建议采用研究专用的数据质量(DQ)评估。

一些评估可以采取设置与研究特别相关的数据质量规则的形式。例如，我们可能想要设置如下规则，即涉及感兴趣的暴露记录中至少有 90%详述了暴露时长。

一项标准评估，涉及审查与 ACHILLES 中的研究最相关的概念，例如研究队列定义中指定的概念。所观察代码的比率随时间发生大幅度变化暗示可能存在数据质量问题。本章后面将讨论一些示例。

另一项评估是，对根据该研究设计的队列定义所产生队列的患病率及患病率随时间的变化进行审查，以评估其是否符合基于外部临床知识的预期。例如，一个新药在被引入市场之前应该不存在暴露，而在上市后随着时间的推移暴露可能增加。同样，某类结局的患病率应与已知人群中该情况的患病率一致。如果一项研究是通过一个数据库网络执行的，那么我们可以比较数据库之间的队列患病率。如果在一个数据库中队列患病率很高，而在另一个数据库中却不存在，则可能存在一个数据质量问题。请注意，这种评估与第 16 章所述的临床效度的概念重叠。我们可能在某些数据库中发现某些患病率与预期不符。这不是因为数据质量问题，而是因为我们的队列定义并未真正获取我们感兴趣的健康状态，或者因为这些健康状态在获取不同患者人群的数据库之间的确有所不同。

15.3.1 映射检查

将源代码映射到标准概念时可能引入错误，而这一步在我们的严格控制中。词汇表中的映射是精心制作的，社区成员注意到的映射中的错误会在词汇表问题跟踪器 1 中报告，并在以后的版本中修复。但是，不可能完全手动检查所有映射，因此错误可能仍然存在。在进行研究时，我们建议审查与该研究最相关的那些概念的映射。幸运的是，这很容易实现，因为在 CDM 中，我们不仅存储标准概念，还存储源代码。我们既可以审查与本研究中使用的概念相对应的源代码，也可以查看不涉及的源代码。

查看映射源代码的一种方法是使用 MethodEvaluation R 软件包中的 checkCohortSourceCodes 函数。此函数使用 ATLAS 创建的队列定义作为输入，并且针对队列定义中使用的每个概念集，它会检查那些映射到集合中概念的源代码。

同时，它还会随着时间的推移计算这些代码的患病率，以帮助识别与特定源代码相关的时间问题。图 15.2 中的示例输出显示了一个名为“抑郁症”的概念集的（部分）分解。在所关注的数据库中，该概念集中最普遍的概念是概念 440383（“抑郁症”）。我们看到数据库中的三个源代码映射到这个概念：ICD-9 代码 3.11，以及 ICD-10 代码 F32.8 和 F32.89。在左侧，我们看到该概念作为一个整体先是随着时间逐渐增加，随后急剧下降。如果我们查看各个代码，就会发现这一现象，可以用 ICD-9 代码在下降时停止使用的事实来解释。即使这是与开始使用 ICD-10 码时间的重合，但 ICD-10 码的普及率总和却比 ICD-9 码小得多。该特定示例是由于 ICD-10 代码 F32.9（“重度抑郁症，单发，未指明”）也应映射到该概念。此问题已在词汇表中解决。

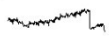
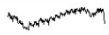
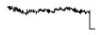
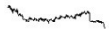
‰ per month	Max monthly ‰	Person count	Description
	26.81	92,019,885	**Depressive Disorder**
	6.64	15,969,198	**Depressive disorder** 440383
	6.64	15,686,275	311 (ICD9CM) Depressive disorder, not elsewhere classified
	0.46	188,230	F328 (ICD10CM) Other depressive episodes
	0.38	94,693	F3289 (ICD10CM) Other specified depressive episodes
	3.10	12,010,783	**Adjustment disorder with mixed emotional features** 433454
	3.07	9,839,712	30928 (ICD9CM) Adjustment disorder with mixed anxiety and depressed mood
	3.03	2,049,618	F4323 (ICD10CM) Adjustment disorder with mixed anxiety and depressed mood
	0.04	121,453	3091 (ICD9CM) Prolonged depressive reaction
	3.17	9,237,192	**Dysthymia** 433440

图 15.2: checkCohortSourceCodes 函数的示例输出

即使前面的示例演示了未映射源代码的一次偶然发现，但通常来说，识别丢失的映射比检查已存在的映射更困难。它要求知道哪些没有映射的源代码是本应映射的。一个执行此评估的半自动方法是，使用 MethodEvaluation R 包中的 findOrphanSourceCodes 函数。此函数允许用户使用简单的文本搜索，在词汇表中搜索源代码，并检查这些源代码是否映射到某一特定概念或其子概念之一。随后将所得的源代码集合限制为仅出现在当前 CDM 数据库中的源代码。例如，在一项研究中，“坏疽性疾病”（439928）及其所有子概念被用来标注所有的坏疽。为了评估它是否确实包括所有指示坏疽的源代码，我们在 CONCEPT 和 SOURCE_TO_CONCEPT_MAP 表中用几个术语（例如“坏疽”）作为描述来检

索，以此标识出源代码。然后使用自动搜索来评估数据中出现的每个坏疽源代码是否确实直接或间接（通过父概念）地映射到“坏疽性疾病”这一概念。评估结果如图 15.3 所示，表明 ICD- 10 代码 J85.0（“坏疽和肺坏死”）仅映射到概念 4324261（“肺坏死”），而后者并不是“坏疽性疾病”的子概念。

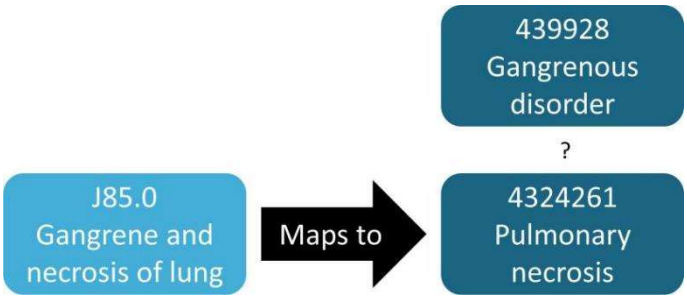


图 15.3: 孤立源代码示例。

15.4 ACHILLES 实践

这里，我们将演示如何针对 CDM 格式的数据库运行 ACHILLES。

我们首先需要告诉 R 如何连接到服务器。ACHILLES 使用 DatabaseConnector 软件包，该软件包提供了一个名为 createConnectionDetails 的函数。键入?createConnectionDetails 得到各种数据库管理系统（DBMS）所需的特定设置。例如，可以使用以下代码连接到 PostgreSQL 数据库：

最后两行定义 cdmDbSchema 变量以及 CDM 版本。稍后我们将使用这些信息告诉 R CDM 格式

```
library(Achilles)
connDetails <- createConnectionDetails(dbms = "postgresql",
                                       server = "localhost/ohdsi",
                                         user = "joe",
                                         password = "supersecret")

cdmDbSchema <-"my_cdm_data" cdmVersion <- "5.3.0"
```

的数据在哪里保存以及使用什么版本的 CDM。请注意，对于 Microsoft SQL Server，数据库架构需要同时指定数据库和架构，例如 cdmDbSchema <- "my_cdm_data.dbo"。

接下来，我们运行 ACHILLES：

```
result <- achilles(connectionDetails, cdmDatabaseSchema = cdmDbSchema,
                                    resultsDatabaseSchema = cdmDbSchema,
                                      sourceName = "My database",
                                      cdmVersion = cdmVersion)
```

此函数将在 resultsDatabaseSchema 中创建多个表，此处我们将这些表设置为与 CDM 数据相同的数据库架构。

我们可以查看 ACHILLES 数据库的特征。可以通过将 ATLAS 指向 ACHILLES 结果数据库，或将 ACHILLES 结果导出到一组 JSON 文件中：

```
exportToJson(connectionDetails,
             cdmDatabaseSchema = cdmDatabaseSchema,
             resultsDatabaseSchema=cdmDatabaseSchema,
             outputPath = "achillesOut")
```

JSON 文件将被写入 achillesOut 子文件夹，并且可以与 AchillesWeb 网络应用程序一起使用提供结果查询功能。例如，图 15.4 显示了 ACHILLES 数据密度图。该图表明，大部分数据始于 2005 年。但是，似乎也有一些大约 1961 年左右的记录，这很可能是数据错误。

另一个示例如图 15.5 所示，揭示了糖尿病诊断代码的患病率激变。这种变化与该特定国家/地区的报销规则变化同时发生，导致出现更多诊断，但可能并不是潜在人群患病率真正增加。

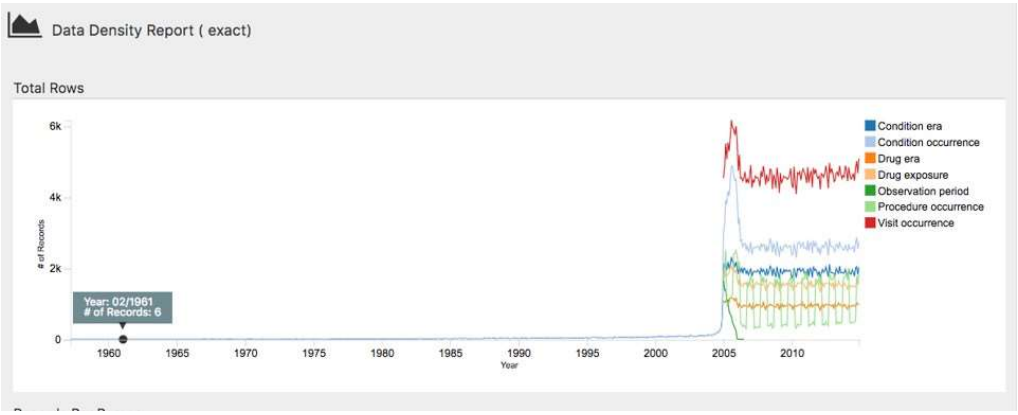


图 15.4： ACHILLES web 浏览器中的数据密度图

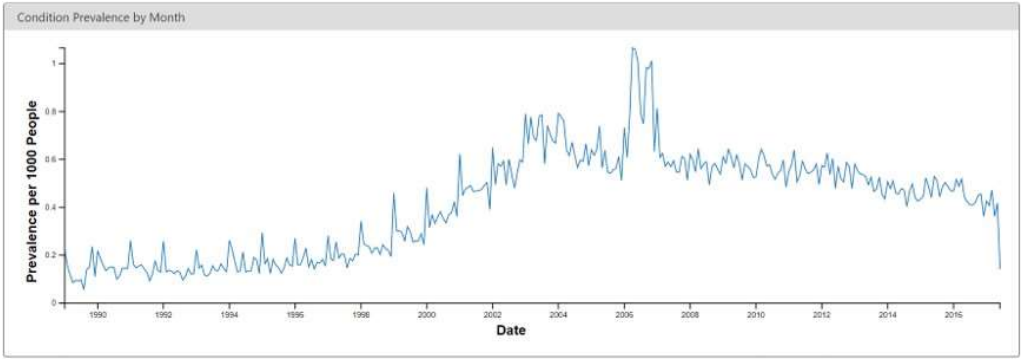


图 15.5: 在 ACHILLES Web 浏览器中的每月的糖尿病编码率

15.5 数据质量控制面板（DQD）实践

在这里，我们将演示如何针对 CDM 格式的数据库运行数据质量控制面板。为此，如 15.4 节中所述，我们对 CDM 连接进行了大量检查。由于目前 DQD 仅支持 CDM v5.3.1，因此，在连接之前请确保您的数据库版本正确。与 ACHILLES 一样，我们需要创建变量 cdmDbSchema 来告诉 R 在哪里寻找数据。

```
cdmDbSchema <- "my_cdm_data.dbo"
```

接下来，我们运行控制面板……

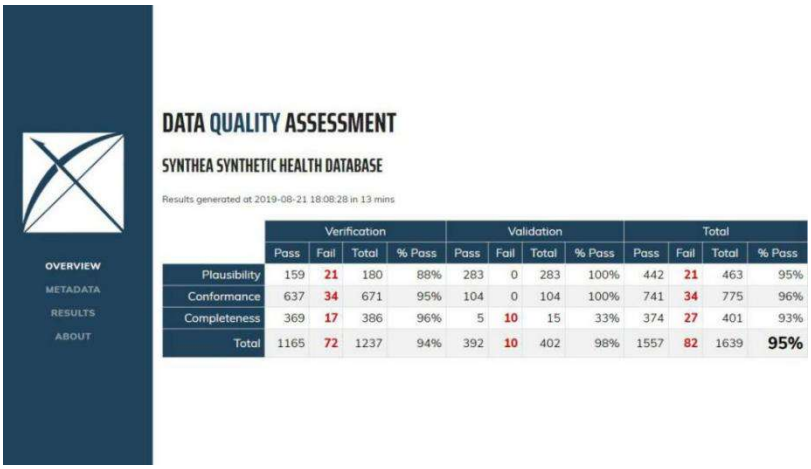
```
DataQualityDashboard::executeDqChecks
                          (connectionDetails = connectionDetails,
                          cdmDatabaseSchema = cdmDbSchema,
                          resultsDatabaseSchema = cdmDbSchema,
                           cdmSourceName = "My database",
                          outputFolder = "My output")

viewDqDashboard(jsonPath)
```

上面的函数将在指定的架构上执行所有可用的数据质量检查。然后，我们向 resultsDatabaseSchema 中写入一个表格，并设置为与 CDM 相同的架构。该表将包含有关每次检查运行的所有信息，包括 CDM 表、CDM 字段、检查名称、检查说明、Kahn 类别和子类别、违反的行数、阈值级别以及检查是否通过等。除了表格外，此函数还将 JSON 文件写入指定为 outputFolder 的位置。使用此 JSON 文件，我们可以通过 Web 浏览器以查看结果。

包含仪表盘结果的变量 jsonPath 应该是 JSON 文件的路径，调用上述 executeDqChecks 函数时产生的 JSON 文件位于 outputFolder。

首次打开控制面板时，将显示概述表，如图 15.6 所示。这里将按内容细分，向您显示每个 Kahn 类别中运行的检查总数、每项通过的次数和百分比以及总通过率。



	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	159	21	180	88%	283	0	283	100%	442	21	463	95%
Conformance	637	34	671	95%	104	0	104	100%	741	34	775	96%
Completeness	369	17	386	96%	5	10	15	33%	374	27	401	93%
Total	1165	72	1237	94%	392	10	402	98%	1557	82	1639	95%

图 15.6: 数据质量控制面板中的数据质量检查概述

点击左侧菜单中的结果，可看到每个已运行检查的深入分析结果（图 15.7）。在此示例表格中显示

的是，检查单个 CDM 表的完整性，或者检查 CDM 中具有至少一条特定表单记录的人员数量和百分比。在该例中，列出的五个表全部为空，控制面板将其视为失败。单击➕图标将打开一个窗口，该窗口将显示产生所列结果所进行的精确查询。这样通过控制面板可以轻松识别被认为失败的行。

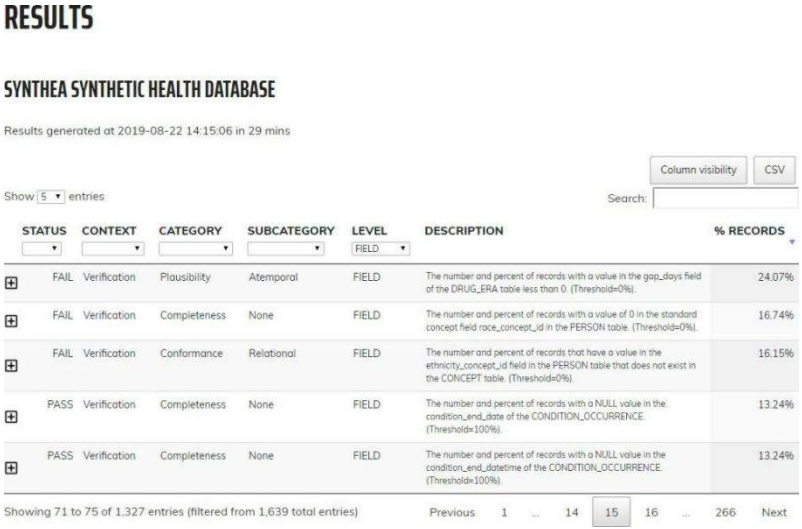


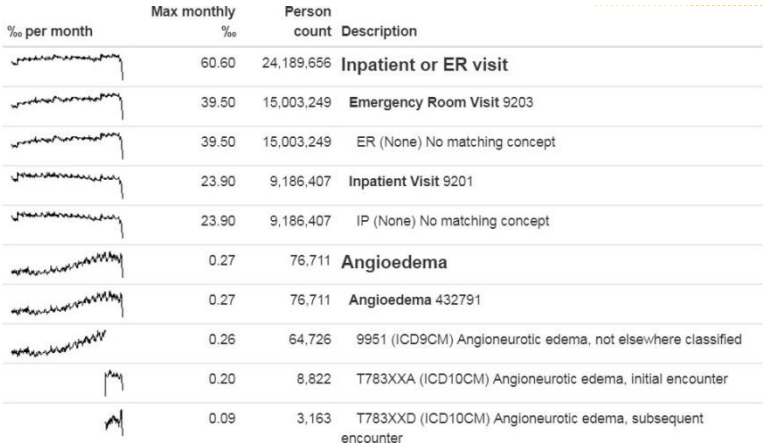
图 15.7: 在数据质量控制面板中进行深入的数据质量检查。

15.6 研究专用质量检查的实践

接下来，我们将针对附录 B.4 中提供的血管性水肿队列定义进行几项检查。我们将假定已按照第 15.4 节中的说明设置了连接的详细信息，并且队列定义 JSON 和 SQL 已分别保存在文件“ cohort.json”和“ cohort.sql”中。可以从 ATLAS 队列定义功能的“导出”选项卡中获取 JSON 和 SQL。

如图 15.8 所示，在 Web 浏览器中打开输出文件。在这里，我们看到血管性水肿队列定义有两个概念集：“住院或急诊就诊”和“血管性水肿”。在此示例数据库中，就诊是通过数据库专用源代码“ ER”和“ IP”找到的，虽然这些词在 ETL 期间已映射到标准概念，但它们不在词汇表中。血管性水肿也是通过一个 ICD-9 和两个 ICD-10 代码找到的。当我们查看单个代码的走势图时，可以清楚地看到两个编码系统之间切换的时间点，但是整个概念集在那个时间点并没有中断。

```
library(MethodEvaluation)
json <- readChar("cohort.json", file.info("cohort.json")$size)
sql <- readChar("cohort.sql", file.info("cohort.sql")$size)
checkCohortSourceCodes(connectionDetails,
                       cdmDatabaseSchema = cdmDbSchema,
                       cohortJson = json,
                       cohortSql = sql,
                       outputFile = "output.html")
```



‰ per month	Max monthly ‰	Person count	Description
	60.60	24,189,656	**Inpatient or ER visit**
	39.50	15,003,249	**Emergency Room Visit** 9203
	39.50	15,003,249	ER (None) No matching concept
	23.90	9,186,407	**Inpatient Visit** 9201
	23.90	9,186,407	IP (None) No matching concept
	0.27	76,711	**Angioedema**
	0.27	76,711	**Angioedema** 432791
	0.26	64,726	9951 (ICD9CM) Angioneurotic edema, not elsewhere classified
	0.20	8,822	T783XXA (ICD10CM) Angioneurotic edema, initial encounter
	0.09	3,163	T783XXD (ICD10CM) Angioneurotic edema, subsequent encounter

图 15.8: 血管性水肿队列定义中使用的源代码。

接下来可以搜索孤立的源代码，这些源代码没有映射到标准概念代码。此处我们寻找标准概念“血管性水肿”，然后寻找名称中包含“血管性水肿”的任何代码和概念，或者我们提供的任何同义词作为其名称的一部分：

```
orphans <- findOrphanSourceCodes(connectionDetails,
                                 cdmDatabaseSchema = cdmDbSchema,
                                 conceptName = "Angioedema",
                                 conceptSynonyms = c("Angioneurotic edema",
                                                     "Giant hives",
                                                     "Giant urticaria",
                                                     "Periodic
                                                     edema"))
View(orphans)
```

代码	描述	词汇表 Id	总数
T78.3XXS	血管神经性水肿，后遗症	ICD10CM	508
10002425	血管性水肿	MedDRA	0
148774	喉血管神经性水肿	CIEL	0
402383003	特发性荨麻疹和/或血管性水肿	SNOMED	0
232437009	喉血管神经性水肿	SNOMED	0
10002472	血管神经性水肿，未在其他分类	MedDRA	0

可以发现，“血管神经性水肿，后遗症”是数据库中实际使用的唯一一个潜在概念，且不应将其映射到血管性水肿。因此，此次分析未发现任何丢失的代码。

15.7 总结

多数观察性医疗数据并不是以研究为目的进行收集的。

数据质量检查是研究的一个组成部分。必须评估数据质量是否达到可用于研究目的的标准。

我们应该常规依据研究的总体目标进行数据质量评估，在某些特定研究中应进行更为深入的评估。

数据质量的某些方面可以通过大量预定义的规则进行自动评价，例如在质量控制面板中已设定的规则。

其他工具也可用来评估与特定研究相关的代码映射。

15.8 练习

练习前的准备

开始练习前，我们假设已按照第 8.4.5 节中的描述安装了 R、R-Studio 和 Java。还需要 SqlRender、DatabaseConnector、ACHILLES 和 Eunomia 软件包，可以使用以下方法安装：

```
install.packages(c("SqlRender", "DatabaseConnector", "devtools"))
devtools::install_github("ohdsi/Achilles")
devtools:install_github("ohdsi/DataQualityDashboard")
devtools::install_github("ohdsi/Eunomia", ref = "v1.0.0")
```

Eunomia 软件包提供了一个 CDM 架构的仿真数据集，该数据集将在您的本地 R 会话中运行。可以使用以下方法获取连接详细信息：

```
connectionDetails <- Eunomia::getEunomiaConnectionDetails()
```

以 CDM 数据库架构为“主”。

练习 15.1 对 Eunomia 数据库执行 ACHILLES。

练习 15.2 对 Eunomia 数据库执行 DataQualityDashboard。

练习 15.3 提取检查的 DQD 列表。

参考答案可以在附录 E.10 中找到。

参考文献

1. Dasu, Tamraparni, and Theodore Johnson. 2003. *Exploratory Data Mining and Data Cleaning*. Vol. 479. John Wiley & Sons.
2. Defalco, F. J., P. B. Ryan, and M. Soledad Cepeda. 2013. “Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure.” *Health Serv Outcomes Res Methodol* 13 (1): 58–67.
3. Hripcsak, G., M. E. Levine, N. Shang, and P. B. Ryan. 2018. “Effect of vocabulary mapping for conditions on phenotype cohorts.” *J Am Med Inform Assoc* 25 (12): 1618–25.
4. Huser, Vojtech, Frank J. DeFalco, Martijn Schuemie, Patrick B. Ryan, Ning Shang, Mark Velez, Rae Woong Park, et al. 2016. “Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Data Sets.” *EGEMS (Washington, DC)* 4 (1): 1239. https://doi.org/10.13063/2327-9214.1239.
5. Huser, Vojtech, Michael G. Kahn, Jeffrey S. Brown, and Ramkiran Gouripeddi. 2018. “Methods for Examining Data Quality in Healthcare Integrated Data Repositories.” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 23: 628–33.
6. Kahn, Michael G., Tiffany J. Callahan, Juliana Barnard, Alan E. Bauck, Jeff Brown, Bruce N. Davidson, Hossein Estiri, et al. 2016. “A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.” *EGEMS (Washington, DC)* 4 (1): 1244. https://doi.org/10.13063/2327-9214.1244.
7. Lei, Johan van der. 1991. “Use and Abuse of Computer-Stored Medical Records.” *Methods of Information in Medicine* 30 (02): 79–80.
8. Makadia, R., and P. B. Ryan. 2014. “Transforming the Premier Perspective Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model.” *EGEMS (Wash DC)* 2 (1): 1110.
9. Matcho, A., P. Ryan, D. Fife, and C. Reich. 2014. “Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model.” *Drug Saf* 37 (11): 945–59.
10. Roebuck, Kevin. 2012. *Data Quality: High-Impact Strategies-What You Need to Know: Definitions, Adoptions, Impact, Benefits, Maturity, Vendors*. Emereo Publishing.
11. Voss, E. A., Q. Ma, and P. B. Ryan. 2015. “The impact of standardizing the definition of visits on the consistency of multi-database observational health research.” *BMC Med Res Methodol* 15 (March): 13.
12. Voss, E. A., R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, F. J. DeFalco, A. Londhe, V. Zhu, and P. B. Ryan. 2015. “Feasibility and utility of applications of the common data model to multiple, disparate observational health databases.” *J Am Med Inform Assoc* 22 (3): 553–64.
13. Weiskopf, Nicole Gray, and Chunhua Weng. 2013. “Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research.” *Journal of the American Medical Informatics Association: JAMIA* 20 (1): 144–51. https://doi.org/10.1136/amiajnl-2011-000681.

第 16 章　临床有效性

本章负责人：Joel Swerdel, Seng Chan You, Ray Chen 和 Patrick Ryan

The likelihood of transforming matter into energy is something akin to shooting birds in the dark in a country where there are only a few birds. 爱因斯坦 1935

OHDSI 的愿景是“一个观察研究对健康与疾病全面了解的世界”。 回顾性设计提供了一种利用现有数据进行研究的工具，但如第 14 章所述，其有效性可能受到诸多方面的威胁。虽然很难将临床有效性与数据质量和统计方法分开，但在这里我们将集中在三个方面探讨临床有效性：卫生保健数据库的特点，队列验证和证据的普遍性。回顾人口水平估计的示例（第 12 章）。我们试图回答以下问题：“与噻嗪类或类似噻嗪类的利尿剂相比，血管紧张素转换酶抑制剂是否更容易引起血管性水肿？”在该示例中，我们证明了与噻嗪类或类似噻嗪类的利尿剂相比，血管紧张素转换酶抑制剂可引起更多的血管性水肿。本章致力于回答这个问题：“我们进行的分析能在多大程度上符合临床意图？”。

16.1 卫生保健数据库的特点

或许我们发现的是血管紧张素转换酶抑制剂的处方与血管性水肿之间的关系，而不是血管紧张素转换酶抑制剂的实际使用与血管性水肿之间的关系，在上一章（第 15 章）中我们已经讨论过了数据质量，通用数据模型（CDM）中转换数据库的质量难以超过原始数据库的质量。在这一节中，我们将讨论大多数医疗应用数据库的特点。OHDSI 中使用的多数数据库都源自医疗保险数据库或电子健康档案系统（EHR），虽然这两者的数据获取过程不同，但研究均不是其主要目的。从医保报销记录中获取的是用于报销的数据元，主要是医生和医疗费用支付者之间的财务交易过程，证明提供者为患者提供的诊疗服务是充分合理的，以使责任方同意支付保单。从 EHR 记录中获取的是支持临床诊疗和卫生行政业务的数据元，它们通常仅反映特定卫生系统中的提供者认为当前有必要记录的医疗服务，以及可为其预期的后续治疗提供依据的必要信息。这些数据元可能无法体现患者的完整病史，也无法整合来自各卫生系统的数据。

研究者对从病人寻求治疗到反映疗效的数据分析过程的理解，有助于其从观察资料中获得可靠证据。例如，“药物暴露量”可以从多种来源的观察资料中推断出，包括临床医生开出的处方、药房配药的记录、医院程序的行政上报数据或患者自述的用药史。数据的来源会影响我们推断哪些患者使用或未使用药物，以及何时使用和使用时长的可信度。数据获取过程可能会导致暴露量的低估，如未记录免费样品或非处方药；或者暴露量的高估，如患者不按处方配药或未坚持服药。了解暴露和结局确认中的潜在偏倚，并对这些测量误差进行更理想化地量化和调整，可以提高我们从现有数据中得出证据有效性的可信度。

16.2 队列验证

Hripcsak 和 Albers（2017）描述过，“表现型是一种对可观察的、潜在变化的生物体状态的规范，而基因型则不同，其由生物体的基因组成衍生而来。‘表现型’一词可被用于由电子健康档案系统（EHR）数据推断出的患者特征。自信息学诞生以来，研究者们就一直在从结构化数据和叙述性数据两方面对 EHR 进行表现型分析，其目的是基于原始 EHR 数据、报销数据或其他临床相关的数据得出有关目标概

念的结论。表现型算法（即识别或表征表型的算法）可以由领域专家和知识工程师生成，包括知识工程方面的最新研究或通过各种形式的机器学习……去生成新的数据表现形式。”

在这一描述中强调了几个在考虑临床有效性时有用的属性：1）明确表明了我们正在谈论的是可观察的事物（因此有可能从我们的观察资料中获取）；2）包含了表现型规范中的时间概念（因为一个人的状态是可以改变的）；3）区分了表现型（预期意图）与表现型算法（预期意图的实现）之间的不同。

OHDSI 采用了“队列”一词来定义在一段时间内满足一个或多个纳入标准的人群。“队列定义”表示根据观察性数据库实例化队列所需的逻辑学。在这方面，队列定义（或表现型算法）可用来建立队列，该队列旨在代表表现型，即我们关注的观察性临床状态的人群。

大多数类型的观察性分析，包括临床特征、人群水平的效应估计和患者水平的预测，都需要在研究过程中建立一个或多个队列。为评估这些分析产生的证据的有效性，每个队列必须考虑到这一问题：基于队列定义和现有的观察资料纳入队列中的个体，他们在多大程度上准确地反映了那些真正属于这一表现型的人群?

回到人群水平估计的例子（第 12 章），“与噻嗪类或噻嗪类利尿剂相比，血管紧张素转换酶抑制剂是否会导致血管性水肿？”，我们必须定义三个队列：一组是血管紧张素转换酶抑制剂新使用者的目标队列，一组是噻嗪类利尿剂新使用者的对照队列，一组是血管性水肿患者的结局队列。我们对患者的血管紧张素转换酶抑制剂或噻嗪类利尿剂的全部服用情况是否完全掌握，从而可以通过首次观察到的暴露来确定其为“新使用者”，而不必担心之前（但未观察到）的服用情况？我们能否轻易地推断出那些具有血管紧张素转换酶抑制剂药物暴露记录的人确实暴露于该药物，而那些没有药物暴露的人确实未暴露于该药物？确定个体处于“使用血管紧张素转换酶抑制剂”状态的持续时间，是否存在不确定性，能否在推断某人开始用药时使其进入队列或停药时退出队列？具有“血管性水肿”病史的人是否确实经历了与其他类型的皮肤过敏反应不同的皮下快速肿胀？有多少比例的血管性水肿患者接受了医学治疗，从而产生了可用于根据队列定义来识别这些临床病例的观察资料？如何将可能由药物引起的血管性水肿与已知由其他因素（例如食物过敏或病毒感染）引起的血管性水肿区别开来？是否充分掌握了疾病发病情况，以使我们有信心得出暴露状态与结局发生率之间的时间关联？这些问题的答案是临床有效性的核心。

在本章中，我们将讨论验证队列定义的方法。我们首先描述衡量队列定义有效性的指标，接下来，介绍两种评估这些指标的方法：1）通过源记录验证进行临床判断；2）PheValuator，这是一种使用诊断预测建模的半自动化方法。

16.2.1 队列评价指标

确定了研究的队列定义后，就可以评估该定义的有效性。评估有效性的一种常用方法是，将一个已定义的队列中的部分或全部个体与参考的“金标准”进行比较，并以混合矩阵的形式展示结果，该矩阵是一个 2*2 列联表，并根据金标准类别和队列定义中的限制对个体分层。图 16.1 展示了混合矩阵的元素。

		金标准	
		真	假
队列定义	真	真阳性	假阴性
	假	假阴性	真阴性

图 16.1：混合矩阵

可以通过将定义应用于一组人群从而来确定队列定义的真假结果。定义中包含的人被认为健康状况阳性，并标记为“真”，那些未被纳入队列定义的人被认为健康状况阴性，并被标记为“假”。虽然队列定义中规定的个体的健康状况的绝对真实性很难去确定，但可以通过多种方法建立一个参考的金标准，其中的两种方法将在本章后面进行介绍。无论使用哪种方法，这些个体的分类都与队列定义中描述的相同。

除了指定表现型的二元指示错误外，健康状况的时间参数也可能产生错误。例如，虽然队列定义可能正确地将个体标记为所属表现型，但该定义可能会错误地指定个体出现健康状况的日期和时间。该误差会增加使用生存分析结果（如危险比）作为效应度量研究的偏倚。

该过程的下一步是评估金标准与队列定义的一致性。同时被金标准方法和队列定义标记为“真”的个体被称为“真阳性”；被金标准方法标记为“假”且被队列定义标记为“真”的个体被称为“假阳性”，即队列定义错误地将这些个体归类为符合条件，然而他们实际上不符合；被金标准方法和队列定义都标记为“假”的个体被称为“真阴性”；被金标准方法标记为“真”而被队列定义标记为“假”的个体被称为“假阴性”，即队列定义错误地将这些个体归类为不符合条件，而事实上他们确实属于表现型。使用混合矩阵中四个单元格的计数，我们可以量化队列定义对一组人群表现型状态分类的准确性。下面是衡量队列定义效果的标准绩效指标：

1. 队列定义的敏感度——根据队列定义，人群中真正属于表现型的个体有多少比例被正确地判别为具有健康结局？由以下公式判定：

敏感度 = 真阳性 / （真阳性 + 假阴性）

2. 队列定义的特异度——根据队列定义，人群中不属于表现型的个体有多少比例被正确地判别为不具有健康结局？由以下公式判定：

特异度 = 真阴性 / （真阴性 + 假阳性）

3. 队列定义的阳性预测值（PPV）——根据队列定义确定的具有健康状况的人群中，有多少比例实际上确实属于表现型？由以下公式判定：

PPV = 真阳性 / （真阳性 + 假阳性）

4. 队列定义的阴性预测值（NPV）——根据队列定义确定的没有健康状况的人群中，有多少比例实际上不属于该表现型？ 由以下公式判定：

NPV = 真阴性 / （真阴性 + 假阴性）

这些指标的满分是 100%，由于观察资料的性质，满分通常远不是标准。Rubbo 等（2015 年）评估了验证心肌梗死队列定义的研究，在他们调查的 33 项研究中，只有一个数据集中的一个队列定义获得了 PPV 满分。总体而言，33 项研究中有 31 项 PPV≥70%。然而他们还发现，在 33 项研究中，只有 11 项记录了敏感度，5 项记录了特异度。PPV 是敏感度、特异度和患病率的函数，在敏感度和特异度保持不变的情况下，具有不同患病率的数据集会产生不同的 PPV 值。如果没有敏感度和特异度，就不能校正由于队列定义不完善造成的偏倚。此外，健康状况的错误分类可能是差异性的，这意味着队列定义在一组人群中的表现情况与对照组相比有所不同，或者当队列定义在两组人群中的表现情况相似时，则是非差异性的。先前的队列定义验证研究未检测潜在的差异性错误分类，即便这可能导致效果评估存在强烈偏倚。

一旦确定了队列定义的绩效指标，该指标就可用于校准运用这些定义进行研究的结果。理论上针对

这些测量误差估计值的校准研究结果已经得到了很好的证实，但在实践中，由于难以获得表型特征，因此很少考虑这些调整。本节其余部分将介绍确定金标准的方法。

16.3 源记录验证

验证队列定义的一种常用方法是，通过源记录验证进行临床判断：由一个或多个学识渊博且能对关注的临床症状或特征进行正确分类的领域专家，对病人的记录进行全面审查。病历纪录审核通常遵循以下步骤：

1. 获得当地机构审查委员会（IRB）和/或相关人员的许可，从而开展包括病历纪录审核在内的研究。
2. 使用队列定义生成队列进行评估。如果没有足够的资源去判定整个队列，则对一部分病人抽样进行人工审核。
3. 指定一个或多个临床专业知识丰富的人员审查病人记录。
4. 确定可判断病人的预期临床状况或特征为阳性或阴性的指南。
5. 临床专家对抽样病例的所有可用资料进行审查和判定，对每个人是否属于表现型进行分类。
6. 根据队列定义分类和临床判定分类将病人列入一个混合矩阵中，并从收集的数据中计算出可能的性能特征。

由于评估中的队列定义只能生成被认为具有预期症状或特征的病人，病历纪录审核的结果通常仅限于对一种性能特征，即阳性预测值（PPV）的评估。因此，根据临床判断，队列样本中的每个人都分为真阳性或假阳性。如果不了解整个人群中所有表现型的患者（包括那些未被队列定义判别的人），就不能识别出假阴性，从而无法补充混合矩阵的其余部分，生成其余的性能特征。识别整个人群中所有表现型患者的潜在方法包括对整个数据库进行病历纪录审核，除非总人群较小，否则这通常是行不通的；或者利用综合临床登记中已经被标记和裁定的真实病例，例如肿瘤登记系统（参见下面的示例）；或者可以对不符合队列定义的人进行抽样，生成一组预计为阴性结果的子集，然后重复上述病历纪录审核的步骤 3-6，检查这些患者是否确实不具备识别真阴性或假阳性的临床症状或特征。这将可以估计出阴性预测值（NPV），同时，如果对表现型患病率有适当的估计量，则可以估算出敏感度和特异度。

通过源记录验证进行临床判断存在许多局限性。如前所述，病历纪录审核是一个非常耗时又费力的过程，即使只是为了评估单个指标（PPV）。这种局限性严重阻碍了评估整个人群从而导致完整填写混合矩阵的实用性。此外，上述过程中的多个步骤可能会使研究结果产生偏差。例如，如果在 EHR 中无法同等地访问所有记录，如果没有 EHR，或者如果需要患者个人同意，那么被评估的子集可能不是真正随机的，并且可能引入抽样或选择偏倚。此外，人工判别容易受到人为误差或错误分类的影响，因此可能无法代表完全准确的衡量标准。由于个人记录中的数据模糊、主观或质量低，临床评判员之间经常会出现分歧。在许多研究中，该过程遵循少数服从多数原则，对患者进行了二分类，这并没有反映出评估者之间的不一致。

16.3.1 源记录验证的示例

哥伦比亚大学欧文医学中心（CUIMC）的一项研究提供了使用病历纪录审核进行队列定义验证的过程示例，该研究验证了多种癌症的队列定义，这也是美国癌症研究所（NCI）可行性研究的一部分。

以这些癌症之一的前列腺癌为例进行验证的步骤如下：

1. 提交 OHDSI 癌症表型研究项目书并获得 IRB 批准。

2. 制定前列腺癌的队列定义：使用 ATHENA 和 ATLAS 扩展词汇创建一个队列定义，包括所有患有前列腺恶性肿瘤 (概念编号 4163261) 的患者，但不包括继发性前列腺肿瘤 (概念编号 4314337) 或前列腺非霍奇金淋巴瘤（概念编号 4048666）。

3. 使用 ATLAS 生成队列，并随机选择 100 名患者进行人工审核，使用对照表将每个 PERSON_ID 映射到患者 MRN。选择 100 名患者的目的是达到我们期望的 PPV 绩效指标的统计精度水平。

4. 人工审核包括住院患者和门诊患者在内的各种 EHR 记录，目的是确定随机子集中的每个个体是否是真或假阳性。

5. 由一名医生进行人工审核和临床判断（尽管在理想情况下，未来更严格的验证研究将会由更多的审查者进行，以评估一致性和评估者间可信度）。

6. 根据完整记录在可用电子病历中的临床文件、病理报告、实验室检测、药物和规程，确定参考标准。

7. 患者被标记为 1）前列腺癌 2）无前列腺癌 3）无法确定。

8. 使用以下公式计算 PPV 的保守估计值：前列腺癌/（无前列腺癌+无法确定）。

9. 接下来，将肿瘤登记作为一个附加的金标准来确定整个 CUIMC 人群的参考标准，计算肿瘤登记系统中被队列定义准确识别和未准确识别的人数，从而能使用这些结果计算真阳性和假阴性个数进而估计敏感度。

10. 利用估计的敏感度、PPV 和患病率，我们可以估算该队列定义的特异度。如前所述，此过程既耗时又费力，因为每个队列定义都必须通过人工病历纪录审核进行单独评估，并与 CUIMC 肿瘤登记系统相关联，以确定所有的绩效指标。尽管在获得肿瘤登记系统的访问权限时进行了快速审核，但 IRB 审批流程和人工病历纪录审核过程仍均需数周时间。

Rubbo 等（2015）对心肌梗死（MI）队列定义验证工作的评估中发现，研究中使用的队列定义以及验证方法和报告的结果之间存在显著的差异。作者得出的结论是，对于急性心肌梗塞没有可用的金标准队列定义。他们指出，这一过程既费钱又费时，由于这一限制，大多数研究的验证样本量较少，导致性能特征估计值存在较大差异。他们还注意到，在 33 项研究中，虽然所有研究都报告了阳性预测值，但只有 11 项研究报告了敏感度，5 项研究报告了特异度。如前所述，如果没有敏感度和特异度的估计值，就无法对错误分类偏倚进行统计学校正。

16.4 PheValuator

OHDSI 社区利用诊断预测模型（Swerdel et al., 2019）开发了一种不同的构建金标准的方法。总体思路是模拟临床医生从原始医疗记录中确定结局指标的方法，但可以利用自动化使其批量运行。该工具已开发为开源的 R 语言包 PheValuator。1PheValuator 使用了 Patient Level Prediction 包中的函数。

具体过程如下所示：

1. 创建一个特异性极高的（“ xSpec”）队列：确定一组人，这些人有着非常高的可能性在训练诊断预测模型时将感兴趣的结局指标用作阳性噪声标签。

2. 创建一个敏感性极高的（“ xSens”）队列：确定一组尽可能多的包括可能有阳性结果的人。该队列将用于识别其相反的情况：即均没有结局指标的一组人，在训练诊断预测模型时用作阴性噪声标签。
3. 使用 xSpec 和 xSens 队列拟合预测模型：如第 13 章所述，我们使用各种可能的患者特征作为预测因子用来拟合模型，并旨在预测某人是否属于 xSpec 队列（有结局指标）或 xSens 队列的对应队列（没有结局指标）。
4. 应用拟合模型来估计一组将被用来评估队列定义表现个体的结局指标概率：该模型的预测变量集可以应用于某人的数据以估计其属于这一表型的预测概率。我们将这些预测称为概率金标准。
5. 评估队列定义的性能特征：我们将预测概率与队列定义的二元分类（混淆矩阵的测试条件）进行比较。使用测试条件和真实条件的估计值，我们可以填充混淆矩阵并估计整个性能特征集，即敏感性，特异性和预测值等指标。

使用这种方法的主要不足是，具有健康结局指标的人的概率估计受到数据库中数据的限制。由于数据库不同，可能无法获得诸如临床病程等重要信息。

在诊断性预测建模中，我们创建一个模型来区分那些患有该疾病的人和那些没有该疾病的人。如患者水平预测章节（第 13 章）中所述，使用目标队列和结局队列来开发预测模型。目标队列包括有或没有健康结局指标的人群；而结局队列则从目标队列中识别那些具有健康结局指标的人。对于 PheValuator 流程而言，我们使用一个特异性极高的队列，即 “ xSpec” 队列，来确定预测模型的结局队列。 xSpec 队列使用了定义来识别那些极有可能有着目标疾病的人。xSpec 队列可以定义为对所关注的健康结局指标存在多次重复出现记录的人。例如，对于房颤，我们可将拥有 10 条或 10 条以上房颤诊断代码记录的人识别为 xSpec 队列。对于心肌梗死，一种急性的结局指标，我们可使用出现 5 次心肌梗死，并且要求至少两次发生在住院期间。预测模型的目标队列是由较低可能性拥有目标健康结局指标的人群与 xSpec 队列中的人群共同构成的。为了确定那些不太可能具有所关注的健康结局指标的人，我们从整个数据库中进行抽样，并排除那些具有某些证据表明属于该表型的人，通常是通过删除记录中具有定义 xSpec 队列概念的个体来进行。但是此方法有一定的局限性。在 xSpec 队列中的人群可能具有与其他有着该疾病人群不同的特征。或者，这部分人与普通病人相比，可能在最初诊断后有着更长的观察期。我们使用 LASSO 逻辑回归来创建用于生成概率金标准的预测模型（Suchard et al., 2013）。该算法生成一个简约模型，通常会删除可能存在于整个数据集中的许多共线性协变量。在当前版本的 PheValuator 软件中，将基于一个人的所有数据（所有观察时间）来评估结局指标状态（是/否），并且不评估队列开始日期的准确性。

16.4.1 利用 PheValuator 进行验证的示例

在需要确定是否已经患有急性心肌梗死患者的研究中，我们可能会使用 PheValuator 来对队列定义的完整性能特征进行评价。

以下是使用 PheValuator 评价心肌梗死队列定义的步骤：

步骤 1：定义 xSpec 队列

确定那些心肌梗死的可能性很高的人。我们要求，住院就诊病历中 5 天内出现了一次或多次任意心肌梗死或其子概念，且在 365 天内记录了 4 次或以上的患者。图 16.2 展示了在 ATLAS 中 MI 的这一队列定义。

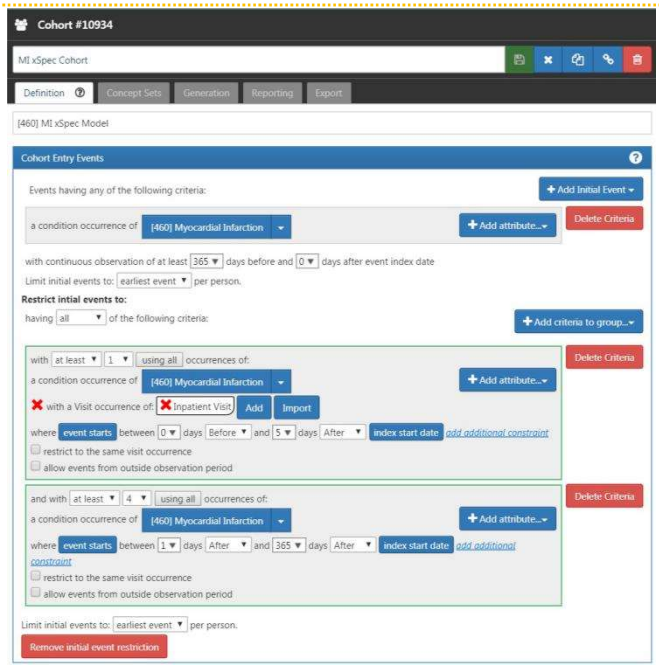


图 16.2: 一个针对心肌梗死特异性极高的队列定义 （xSpec）。

步骤 2: 定义 xSens 队列

我们又建立了一个敏感性极高的队列（xSens）。对于心肌梗死，可以将该队列定义为在其病史中任何时候具有至少一次包含心肌梗死概念的记录的患者。图 16.3 展示了在 ATLAS 中心肌梗死的 xSens 队列定义。

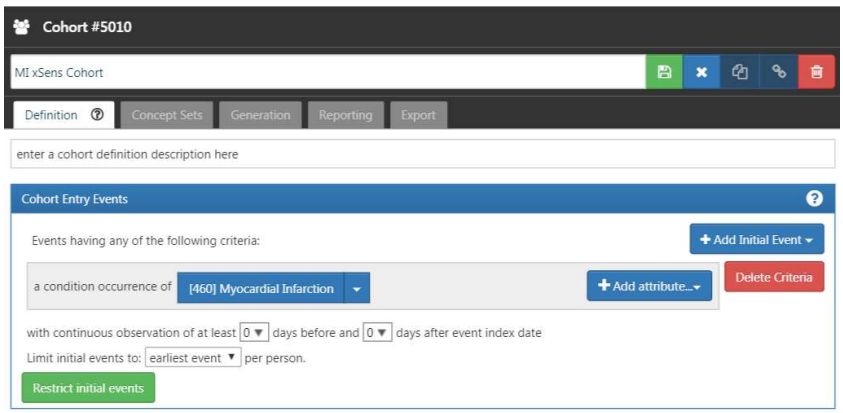


图 16.3: 一个针对心肌梗死敏感性极高的队列定义 （xSens）。

步骤 3: 拟合预测模型

函数 createPhenoModel 建立了诊断预测模型，用于估算评估队列中具有目标健康结局指标的概率。要使用此函数，我们利用在步骤 1 和 2 中开发的 xSpec 和 xSens 队列。在函数中将 xSpec 队列以 xSpecCohort 参数输入，将 xSens 队列以 exclCohort 参数输入，以指示应将 xSens 队列排除在

建模过程中使用的目标队列之外。使用这种排除方法，我们可以确定有较低可能性具备目标健康结局指标的患者。我们可以将这一组视为“阴性噪声”患者，即他们有较低可能性存在目标健康结局指标，但允许包含少量阳性健康结局的人。我们还可以利用函数中的 prevCohort 参数引入 xSens 队列。在此过程中使用此参数来确定人群中具有健康结局指标的大致患病率。通常，从数据库中随机抽取大量样本，应产生这样一种人群，其中具有目标结局指标的人数大约与数据库中的患病率成比例。使用我们描述的方法，我们不再拥有随机样本，因此需要将有无结局指标的人数按比例重新设置为基准，并重新校准预测模型。

建模过程中必须排除用于定义 xSpec 队列的所有概念。为此，我们将 excludedConcepts 参数设置为 xSpec 定义中使用的概念列表。例如，对于 MI，我们使用心肌梗死及其所有子概念，并在 ATLAS 中创建了一个概念集。在此示例中，我们将 excludedConcepts 参数设置为 4329847（心肌梗死的概念 ID），还将 addDescendantsToExclude 参数设置为 TRUE，表示所有排除概念的子概念也应被排除。

有几个参数可用于指定建模过程中所包含人员的特征。我们可以通过将 lowerAgeLimit 设置为模型中期望的年龄下限，将 upperAgeLimit 设置为上限，来设置建模过程中所包含人员的年龄。尤其是在为特定年龄段的人群制定计划研究的队列定义时，我们不妨这样做。例如，如果一项研究中使用的队列定义是针对儿童的 1 型糖尿病，则您可能希望将用于开发诊断预测模型的年龄限制为 5 至 17 岁。于是，我们将生成一个模型，其功能可能与要检验的队列定义所选择的人员更为紧密相关。我们还可以通过将 gender 参数设置为男性或女性的概念 ID 来指定模型中包括哪种性别。默认情况下，该参数设置同时包括男性和女性。此功能在特定性别的健康结局指标（例如前列腺癌）中会用到。我们可以通过将 startDate 和 endDate 参数分别设置为日期范围的上限和下限，根据个人记录中的首次诊断来设置包含个人的时间范围。最后，mainPopnCohort 参数可用于指定一个大的队列，从中选择目标队列和结局队列中的所有人。在大多数情况下，该值将设置为 0，表示在选择目标人群和结果人群时没有限制。但是，有时此参数可用于构建更好的模型，可能是在健康结果的患病率极低（也许为 0.01%或更低）的情况下。例如：

```
setwd("c:/temp")
library(PheValuator)
connectionDetails                  <-
  createConnectionDetails(   dbms    =
  "postgresql",
  server         =
  "localhost/ohds
  i",    user    =
  "joe",
  password = "supersecret")


phenoTest                  <-
  createPhenoModel(  connecti
  onDetails                 =
  connectionDetails,
  xSpecCohort     =     10934,
  cdmDatabaseSchema         =
  "my_cdm_data",
  cohortDatabaseSchema      =
  "my_results",
  cohortDatabaseTable       =
  "cohort",
  outDatabaseSchema = "scratch.dbo", #应有写权限
  trainOutFile               =
  "5XMI_train",  exclCohort =
  1770120, #the xSens cohort
  prevCohort = 1770119, #决定患病率的队列
  modelAnalysisId         =
  "20181206V1",
  excludedConcepts          =
  c(312327,         314666),
  addDescendantsToExclude
  =  TRUE,  cdmShortName  =
  "myCDM",
  mainPopnCohort = 0, #使用整个群体
  lowerAgeLimit = 18,
  upperAgeLi
  mit  =  90,
  gender    =
  c(8507,
  8532),
  startDate  =
  "20100101",
  endDate = "20171231")
```

在此示例中，我们使用了在“ my_results”数据库中定义的队列，指定了队列表的位置（cohortDatabaseSchema，cohortDatabaseTable-“ my_results.cohort”）以及模型将在其中找到条件，药物暴露量等的位置（cdm- DatabaseSchema-“ my_cdm_data”）。 该模型中包含的是在 2010 年 1 月 1 日至 2017 年 12 月 31 日之间首次访问 CDM 的患者。

我们还专门排除了用于创建 xSpec 队列的概念 ID 312327、314666 及其子概念。首次访视时他们的年龄在 18 至 90 岁。使用上述参数，此步骤输出的预测模型的名称将为：“ c:/temp/lr_results_5XMI_train_myCDM_ePPV0.75_20181206V1.rds”。

步骤 4: 建立评价队列

createEvalCohort 函数使用 PatientLevelPrediction 包中的 applyModel 函数来产生大样本队列，每个个体有目标健康结局指标的预测概率。该函数需要指定 xSpec 队列（通过将 xSpecCohort 参数设置为 xSpec 队列 ID）。我们还可以像上一步一样指定评估队列中人员的特征。这可以包括指定年龄上限和下限（分别通过设置年龄，分

```
setwd("c:/temp")
connectionDetails                    <-
  createConnectionDetails(     dbms      =
  "postgresql",
  server             =
  "localhost/ohdsi",
  user = "joe",
  password = "supersecret")

        evalCohort <-
  createEvalCohort( connectionDe
     tails=connectionDetails,
      xSpecCohort = 10934,
       cdmDatabaseSchema=
          "my_cdm_data",
      cohortDatabaseSchema=
          "my_results",
     cohortDatabaseTable =
  "cohort", outDatabaseSchema=
  "scratch.dbo", testOutFile =
   "5XMI_eval", trainOutFile =
  "5XMI_train", modelAnalysisId
        = "20181206V1",
       evalAnalysisId =
  "20181206V1", cdmShortName =
   "myCDM", mainPopnCohort = 0,
  lowerAgeLimit =18,
  upperAgeLimit
  =  90,  gender =
  c(8507,   8532),
  startDate      =
  "20100101",
  endDate = "20171231")
```

别设定 lowerAgeLimit 和 upperAgeLimit 参数），性别（通过将 gender 参数设置为男性和/或女性的概念 ID），开始和结束日期（分别通过将 startDate 和 endDate 参数设置为日期），并通过设置 mainPopnCohort 为该队列要使用的队列 ID 群。例如：

```
  startDate      =
  "20100101",
  endDate = "20171231")
```

在此示例中，参数指定函数应使用模型文件：“ c：/temp/lr_results_5XMI_train_myCDM_ePPV0.75_20181206V1.rds”以生成评估队列文件：

“ c：/temp/lr_results_5XMI_eval_myCDM_ePPV0.751”。 模型和在此步骤中创建的评估队列文件将用于下一步提供的队列定义的评估之中。

步骤 5：创建并测试队列定义

下一步是创建并测试要评估的队列定义。期望的性能特征应取决于该队列用于解决感兴趣的研究问题的预期用途。对于某些特定的问题，可能需要一个非常敏感的算法，而其他情况可能需要更特异的算法。图 16.4 显示了使用 PheValuator 确定队列定义的性能特征的过程。

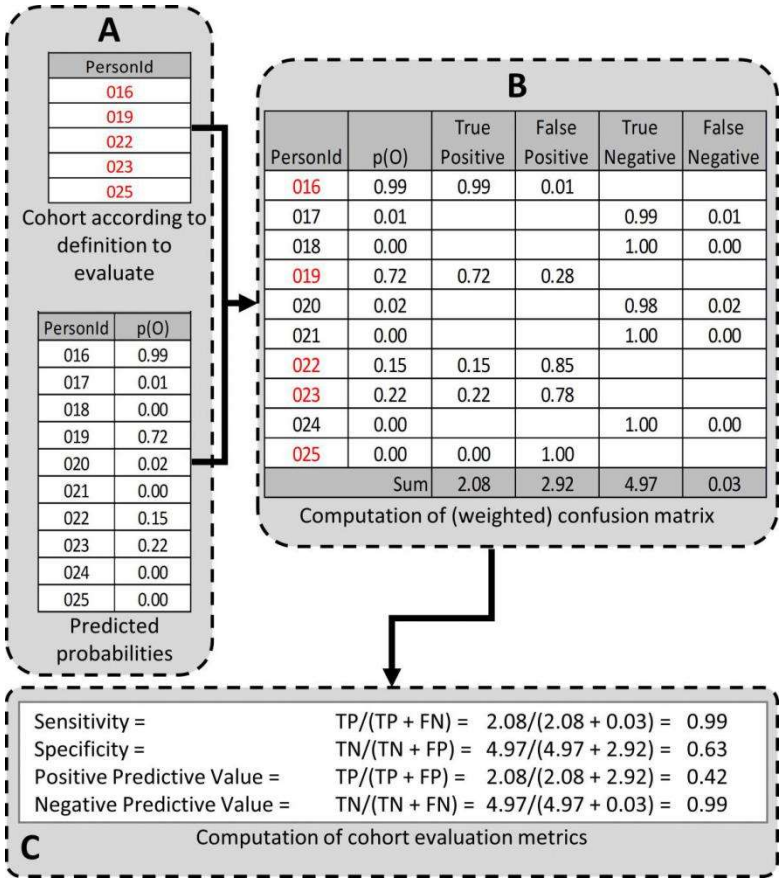


图 16.4: 使用 PheValuator 确定队列定义的性能特征。p(O) = 结局指标的概率；TP = 真阳性； FN = 假阴性；TN = 真阴性；FP = 假阳性。

在图 16.4 的 A 部分中，我们检查了要测试的队列定义中的个体，并从评估队列中（已在上一步中创建）找到了包含在队列定义中的个体（ID 016、019、022、023 与 025），以及未包括在内的个体（ID 017、018、020、021 和 024）。对于这些包括在内/排除在外的个体，我们先前已经使用预测模型（p(O)）确定了健康结局指标的可能性。

我们按以下方式估计“真阳性”，“真阴性”，“假阳性”和“假阴性”的值（图 16.4 的 B 部分）：

1. 如果队列定义包含评估队列中的某个人，即该评估队列定义将该人视为“阳性”。健康结局指标的预测概率表示该人为“真阳性”贡献的计数的预期值，而 1 减去这一概率则表示对该人的“假阳性”贡献的计数的预期值。我们将所有个体的所有期望值相加以获得总期望值。例如，PersonId 016 预测存在健康结局指标的概率为 99%，将真阳性增加 0.99（计数的期望值增加 0.99），并将假阳性增加 1.00-0.99 = 0.01。对队列定义中包括的评估队列中的所有个体（即 PersonIds 019、022、023 和 025）重复这一操作。

2. 同样的，如果队列定义中未包含评估队列中的某个人，即队列定义将某人视为“阴性”，1 减

去该人表型的预测概率即归因于真阴性的计数的期望值，并将其累加，与此同时，将预测概率表型的能力是计数的预期值归因于假阴性，并将其累加到假阴性中。例如，PersonId 017 预测健康结局指标存在的可能性为 1%（相应地，无健康结局指标的可能性为 99%），并且将 1.00 – 0.01 = 0.99 添加到真阴性中，并将 0.01 加入假阴性中。对评估队列中未包含在队列定义中的所有个体（即 PersonIds 018、020、021 和 024）重复此操作。

对评估队列中的所有个体进行如上计算之后，我们将每个单元格计数的期望值填充到混淆矩阵的四个单元格中，就可以创建 PA 性能特征的点估计值，例如敏感性，特异性和阳性预测值（图 1C）。在此我们强调，这些预期的单元格计数不能用于评估估计的方差，只能用于评估点估计值。在该示例中，敏感性，特异性，PPV 和 NPV 分别为 0.99、0.63、0.42 和 0.99。

testPhenotype 函数用来确定队列定义的性能特征。此函数使用我们在前两个步骤中建立的模型和评估队列的输出结果。我们将 modelFileName 参数设置为 createPhenoModel 函数输出的 RDS 文件在本示例中为：

“ c/temp/lr_results_5XMI_train_myCDM_ePPV0.75_20181206V1.rds ”。我们将 resultFileName 参数设置为从 createEvalCohort 函数输出的 RDS 文件，在本示例中为“ c/temp/lr_results_5XMI_eval_myCDM_ePPV0.75_20181206V1.rds”。为了测试我们希望在研究中使用的队列定义，我们将 cohortPheno 设置为该队列定义的队列 ID。我们可以将 phenText 参数设置为该队列定义的人类可读语言描述，例如“ 院内场景下心肌梗死的发生率”。我们将 testText 参数设置为 xSpec 定义的可读描述，例如“ 5 次心肌梗死”。此步骤的输出是一个数据框，其中包含测试队列定义的性能特征。 cutPoints 参数是一个数据列表，用于得出性能特征结果值。通常使用图 1 中所述的“期望值”来计算性能特征。我们在 cutPoints 参数的列表中包含了“ EV” ， 以便基于期望值检索性能特征。我们可能还希望根据特定的预测概率（即分界值）查看性能特征。例如，如果希望获得健康结局为阳性的预期概率等于或高于 0.5，而健康结局为阴性的预期概率低于 0.5 的所有情况的性能特征，那么需将“0.5”添加 cutPoints 参数列表。比如：

```
  setwd("c:/temp")

connectionDetails <-
  createConnectionDe
    tails( dbms =
    "postgresql",

      server =
  "localhost/ohdsi",
    user = "joe",

     password =
    "supersecret")

   phenoResult <-
  testPhenotype( con
   nectionDetails =
  connectionDetails,

 cutPoints = c(0.1,
    0.2, 0.3, 0.4,
   0.5, "EV", 0.6,
   0.7, 0.8, 0.9),

  resultsFileName =
  "c:/temp/lr_result
  s_5XMI_eval_myCDM_
  ePPV0.75_20181206V
       1.rds",

   modelFileName =
  "c:/temp/lr_result
  s_5XMI_train_myCDM
  _ePPV0.75_20181206
       V1.rds",

    cohortPheno =
       1769702,

 phenText = "All MI
   by Phenotype 1 X
    In-patient, 1st
  Position", order =
          1,

testText = "MI xSpec
   Model - 5 X MI",
  cohortDatabaseSche
  ma = "my_results",
     cohortTable =
       "cohort",

   cdmShortName =
       "myCDM")
```

在该示例中，提供了包括预期值（“EV”）在内的大范围的预测阈值（分界值）。已知给定该参数设置，此步骤的输出将提供每个预测阈值以及使用预期值计算的性能特征（即灵敏度，特异性等）。评估将预测信息用于先前步骤中建立的评估队列。可以将此步骤产生的数据框保存到 csv 文件中以便详细查看。使用这一流程，表 16.1 展示了五个数据集中心肌梗死的四个队列定义的性能特征。对于类似于

Cutrona 及其同事评估的队列定义，“>=1 X HOI，住院患者”，我们发现平均 PPV 为 67%（范围：59%-74%）。

表 16.1: 利用 pheValuator 使用诊断条件编码在多个数据集上确定心肌梗死的四个队列定义的性能特征。Sens – 敏感性; PPV – 阳性预测值 ; Spec – 特异性; NPV – 阴性预测值; Dx Code – 队列的诊断编码。

表型算法(Phenotype Algorithm)	数据库 (Database)	Sens	PPV	Spec	NPV
>=1 X HOI	CCAE	0.761	0.598	0.997	0.999
	Optum1862	0.723	0.530	0.995	0.998
	OptumGE66	0.643	0.534	0.973	0.982
	MDCD	0.676	0.468	0.990	0.996
	MDCR	0.665	0.553	0.977	0.985
>= 2 X HOI	CCAE	0.585	0.769	0.999	0.998
	Optum1862	0.495	0.693	0.998	0.996
	OptumGE66	0.382	0.644	0.990	0.971
	MDCD	0.454	0.628	0.996	0.993
	MDCR	0.418	0.674	0.991	0.975
>=1 X HOI, In-Patient	CCAE	0.674	0.737	0.999	0.998
	Optum1862	0.623	0.693	0.998	0.997
	OptumGE66	0.521	0.655	0.987	0.977
	MDCD	0.573	0.593	0.995	0.994
	MDCR	0.544	0.649	0.987	0.980
1 X HOI, In-Patient, 1st Position	CCAE	0.633	0.788	0.999	0.998
	Optum1862	0.581	0.754	0.999	0.997
	OptumGE66	0.445	0.711	0.991	0.974
	MDCD	0.499	0.666	0.997	0.993
	MDCR	0.445	0.711	0.991	0.974

16.5 证据的泛化

尽管可以在给定的观察性数据库的背景下对队列进行很好的定义和全面评估，但临床有效性依然会受到结果被认为泛化到目标人群不同程度的限制。对同一主题的多次观察性研究可能会产生不同的结果，这不仅可能归咎于其设计和分析方法不同，还可能是由于数据源的选择所致。Madigan 等(2013b)证明了数据库的选择会影响观察性研究的结果。他们系统地研究了 10 个观察性数据库中 53 种药物治

疗配对和两种研究设计（队列研究和自身对照的病例系列研究）结果的异质性。 即使他们保持研究设计不变，观察到的效果估计也存在很大的异质性。

在 OHDSI 网络中，观察性数据库在其所代表的人群（如儿科与老年人，私人保险雇员与公共保险失业者），采集数据的医疗环境（如住院病人与门诊病人，基础医疗与二级/专科医疗），数据收集过程（如行政上报数据，EHR，临床登记注册），以及作为医疗基础的国家和地区卫生体系等方面，都存在着巨大差异。这些差异可能表现为在研究疾病和医疗干预效果时观察到的异质性，也可能影响将为网络研究提供证据的每个数据源质量的可信度。尽管 OHDSI 研究网络中的所有数据库均已按照 CDM 进行了标准化，但必须强调的是，标准化并不会减少整个人群中存在的真正的固有异质性，只是提供了一致的框架来调查和更好地理解跨网络的异质性。OHDSI 研究网络提供了一个环境，可将相同的分析过程应用到世界各地的各种数据库中，以便研究人员可以在保持其他方法方面不变的情况下，解释跨多个数据源的结果。 OHDSI 在网络研究中采用开放科学的协作方法，让参与数据研究的伙伴与具有临床领域知识的研究人员以及具有分析专业知识的方法学家共同合作，可以使整个研究网络中对数据的临床有效性的理解达成某种共识，为使用这些数据作为证据建立可信度提供了基础。

16.6 总结

- 通过了解基础数据源的特征，评估分析中队列的表现特征以及评估研究对目标人群的泛化能力，来建立临床有效性。
- 可以根据具体的定义对队列定义进行评估，并基于队列定义和现有观察数据准确反映出真正属于该表型的个体。
- 队列定义验证需要估计多个性能特征，包括敏感性，特异性和阳性预测值，以全面总结并校正测量误差。
- 通过源记录验证和PheValuator进行临床判定，是评估队列定义验证的两种代表性方法。

参考文献

1. Hripcsak, G., and D. J. Albers. 2017. “High-fidelity phenotyping: richness and freedom from bias.” J Am Med Inform Assoc, October.
2. Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. 2013. “Evaluating the impact of database heterogeneity on observational study results.” Am. J. Epidemiol. 178 (4): 645–51.
3. Rubbo, B., N. K. Fitzpatrick, S. Denaxas, M. Daskalopoulou, N. Yu, R. S. Patel, H. Hemingway, et al. 2015. “Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations.” Int. J. Cardiol. 187: 705–11.
4. Suchard, M. A., S. E. Simpson, Ivan Zorych, P. B. Ryan, and David Madigan. 2013. “Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model.” ACM Trans. Model. Comput. Simul. 23 (1): 10:1–

10:17. https://doi.org/10.1145/2414416.2414791.

5. Swerdel, J. N., G. Hripcsak, and P. B. Ryan. 2019. "PheValuator: Development and Evaluation of a Phenotype Algorithm Evaluator." J Biomed Inform, July, 103258.

58. https://github.com/OHDSI/PheValuator

第 17 章　软件有效性

章节负责人：马丁·舒米（Martijn Schuemie）

软件有效性的中心问题是：

该软件可以完成预期的工作吗？

软件有效性是证据质量的重要组成部分：只有我们的分析软件能够完成预期的工作，我们才能提供可靠的证据。如第 17.1.1 节所述，将每项研究视为一项软件开发工作，创建执行从通用数据模型（CDM）中的数据到结果（如估值和图表）整个分析过程的自动化脚本，必须验证此脚本以及此脚本中使用的所有软件的有效性。如 8.1 节所述，我们可以将整个分析编写为自定义代码，也可以使用 OHDSI 方法库中的可用函数。使用方法库的优点在于，其有效性已经得到最大限度确保，因此确定整个分析的有效性变得不那么麻烦。

在本章中，我们首先描述编写有效分析代码的最佳方案。之后，我们讨论如何通过其软件开发和测试来验证方法库。

17.1 代码有效性研究

17.1.1 自动化是可重复性的要求

传统上，观察性研究通常被视为一个旅程而不是一个过程：数据库专家从数据库中提取数据集并将其交给数据分析师，后者可以在电子表格编辑器或其他交互式工具中打开它，然后开始进行数据分析。分析最终产生了结果，但对结果如何产生的却鲜有保留。旅程的目的地已到达，但无法追溯到达目的地的确切步骤。这种做法是完全不能接受的，因为它不仅不可重复，而且缺乏透明度。因为我们不知道结果到底是如何产生的，所以我们也不能确定其中是否出错。

因此，每次产生证据的分析都必须完全自动化。自动化是指分析应作为单个脚本实施，且能用一个命令复现从 CDM 格式的数据库到图表结果的整个分析。分析可以是任意复杂度的，可以只产生一个计数，也可以是为数百万个研究问题生成根据经验校准的估值，但是适用相同的原则。脚本可以调用其他脚本，而其他脚本又可以调用更低级别的分析过程。

尽管在 OHDSI 中首选语言是 R 语言，但你可以使用任何计算机语言来实现分析脚本。得益于 DatabaseConnector R 软件包，我们可以直接连接到 CDM 格式的数据，并且许多高级分析都可以通过 OHDSI 方法库中的其他 R 语言软件包实现。

17.1.2 编程最佳方案

观察性分析可能会非常复杂，需要很多步骤才能得出最终结果。这种复杂性会使维护分析代码变得更加困难，并且增加了出错的可能性，同时也使发现错误变得更加困难。幸运的是，多年来，计算机程序员在处理复杂性的代码时已开发出最佳方案。这些方案易于阅读，可重复使用，方便改编和利于验证。对这些最佳方案的完整讨论很长，在这里，我们重点介绍这四个重要原则：

1. 抽象：我们可以以“函数”为单位组织代码，而不是编写一个可以完成所有任务的大型脚本，这会导致所谓的“面条式代码”，即代码行之间的依赖关系可能从任何地方到任何地方（例如，

在第 1000 行中使用在第 10 行设置的值)。函数应该有一个明确的目标，例如“随机抽样”，一旦创建我们就可以在较大的脚本中使用此函数，而不必考虑函数的细节。我们可以将函数抽象为一个易于理解的概念。

2. 封装：为了使抽象起作用，我们应确保将函数的依赖性最小化并明确定义。我们的示例采样函数应包含一些自变量（例如，数据集和样本大小）和一个输出（例如，样本)。其他因素都不应该影响函数的功能。应该避免使用所谓的“全局变量”，即在函数外部设置的，不是函数的自变量，但仍在函数中使用的变量。
3. 命名清晰：变量和函数应具有清晰的名称，使代码几乎像自然语言一样易读。例如，我们可以编写如下代码：sampledPatients <- takeSample(patients, sampleSize = 100)，而不是 x <- spl(y, 100)。尽量不要缩写，现代语言对变量和函数名称的长度没有限制。
4. 可重复使用：编写清晰、封装良好的函数的一个优点是它们通常可以被重复使用。这不仅节省了时间还意味着代码量更少，从而复杂性更低，出错的几率更小。

17.1.3 代码验证

有几种方法来验证软件代码的有效性，但是有两种方法对于实施观察性研究的代码经常用到:

1. 代码审查：一个人编写代码，另一个人审查代码。
2. 双重代码： 两个人都独立编写分析代码，然后比较两个脚本的结果。

代码审查的优点是通常工作量少，但缺点是审查者可能会遗漏一些错误。另一方面，双重编码通常会占用大量人力，但是不太可能遗漏错误。双重编码的另一个缺点是，由于要做出许多次要的选择，因此两个单独的实现几乎总是产生不同的结果（例如，是否应将“直到曝光结束”解释为包括曝光结束日期？)。结果，两个本该独立的程序员经常需要一起工作以使他们的分析保持一致，从而破坏了他们的独立性。

其他软件验证方案（例如单元测试）在这里不太相关，因为研究通常是一次性活动，在输入（CDM 中的数据）和输出（研究结果）之间具有高度复杂的关系，从而使这些方案的可用性降低。注意这些做法适用于方法库。

17.1.4 使用方法库

OHDSI 方法库提供了一个大的函数集，允许仅使用几行代码即可实现大多数观察性研究。因此，使用方法库将把确定学习代码的有效性的大部分负担转移给方法库。方法库的软件开发过程和大量测试可确保其有效性。

17.2 方法库软件开发流程

OHDSI 方法库由 OHDSI 社区开发。在两个地方对方法库的修改建议进行讨论：GitHub 问题跟踪器(例如 CohortMethod 问题跟踪器 1)和 OHDSI 论坛 2 。两者均对公众开放。社区的任何成员都可以向该库贡献软件代码，但是，对软件发行版本中包含的任何更改只能由 OHDSI 人群水平估计工作组的领导(Marc Suchard 博士和 Martijn Schuemie 博士)和 OHDSI 患者水平预测工作组的领导(Peter Rijnbeek 博士和 Jenna Reps 博士）做出最终批准。

用户可以直接从 GitHub 仓库的主分支安装 R 语言中的方法库，也可以通过一个名为 drat 的系统

安装，该系统始终与主分支保持最新。R 语言的综合 R 存档网络（CRAN）提供了许多方法库软件包，并且软件包的数目预计会越来越多。

OHDSI 采用合理的软件开发和测试方法来最大限度地提高方法库性能的准确性、可靠性和一致性。重要的是，由于方法库是根据 Apache License V2 的条款发布的，因此方法库下的所有源代码，无论是 R、C++、SQL 还是 Java，都可以供 OHDSI 社区的所有成员和公众进行同行评审。因此，方法库中包含的所有功能都需要对其准确性、可靠性和一致性进行不断的评估和改进。

17.2.1 源代码管理

方法库的所有源代码都通过源代码版本控制系统 git 管理，可以通过 GitHub 公开访问。OHDSI 方法库存储库是受访问限制的。世界上任何人都可以查看源代码，OHDSI 社区的任何成员都可以通过所谓的拉取请求提交更改。只有 OHDSI 人群水平估计工作组和 OHDSI 患者水平预测工作组的领导才能批准此类请求，更改主分支并发布新版本。代码更改的连续日志在 GitHub 存储库中维护，并反映代码和文档更改的所有方面。这些提交日志可供公众查看。

OHDSI 人群水平估计工作组和患者水平预测工作组领导将根据需要发布新版本。新版本发行时首先将更改推送到主分支，更改的软件包的版本号（在包内的描述文件中定义）大于前一个发行版本号。这将自动触发包的检查和测试。如果所有检查和测试都通过，则新版本将在版本控制系统中自动标记，并将软件包将自动上传到 OHDSI drat 存储库。新版本使用三分量版本号来编号:

新的微型版本(例如从 4.3.2 到 4.3.3)只表明错误已经修复。没有新的功能，并保证向前和向后的兼容性

新的次要版本(例如从 4.3.3 到 4.4.0)表明增加了功能。只能保证向后兼容

新的主要版本(例如从 4.4.0 到 5.0.0)表明有主要的修订。不能保证兼容性

17.2.2 文档

方法库中的所有软件包都通过 R 语言的内部文档框架进行了文档记录。每个软件包都有一个说明手册，其中描述了软件包中的每个函数。为了促进函数文档和函数实现之间的一致性，我们可以使用 roxygen2 软件将函数文档和源代码合并到一个文件中。说明手册可以通过 R 语言的命令行界面、软件包存储库中的 PDF 文件和网页按需查看。此外，许多软件包还具有突出特定用例的小插图。所有文档都可以通过方法库网站 3 查看。

终端用户可以使用所有方法库源代码。使用 GitHub 的问题跟踪系统和 OHDSI 论坛可以促进来自社区的反馈。

17.2.3 获取当前和历史存档版本

方法库软件包的当前和历史版本可以在两个位置获取。第一，GitHub 版本控制系统包含每个软件包的完整开发历史，并且可以重构和检索软件包在每个时间点的状态。最重要的是，每个发布的版本都在 GitHub 中标记。第二，已发布的 R 语言源软件包存储在 OHDSI GitHub drat 库中。

17.2.4 维护，支持和停用

OHDSI 在错误报告，修复和补丁方面积极支持方法库当前的每个版本。可以通过 GitHub 的问题

跟踪系统和 OHDSI 论坛报告问题。每个软件包都有一个说明手册，以及一个或多个小插图（有的未提供）。提供在线视频教程，并不时提供面对面教程。

17.2.5 人员资质

OHDSI 社区的成员代表了多个统计学科，并位于多个大洲的学术机构、非营利组织和行业附属机构。

OHDSI 人群水平估计工作组和 OHDSI 患者水平预测工作组的所有领导都拥有来自认可的学术机构的博士学位，并在同行评议期刊上发表了大量论文。

17.2.6 物理和逻辑安全

OHDSI 方法库托管在 GitHub4 系统上。GitHub 的安全措施在 https://github.com/security 上进行了描述。OHDSI 社区的所有成员都需要用户名和密码才能对方法库进行修改，只有人群水平估计工作组和 OHDSI 患者水平预测工作组的领导才能对主分支进行修改。根据标准的安全策略和功能需求，用户的帐户访问受到限制。

17.2.7 灾难恢复

OHDSI 方法库托管在 GitHub 系统上。GitHub 的灾难恢复功能在 https://github.com/security 上进行了描述。.

17.3 方法库测试

我们要区分在方法库上执行的两种类型的测试：包中各个功能的测试（又名“单元测试”），以及使用仿真的更复杂功能的测试。

17.3.1 单元测试

OHDSI 维护并更新了大量的自动验证测试，从而能够针对已知数据和已知结果对源代码进行测试。每次测试都从指定的一些简单的输入数据开始，然后针对该输入执行包中的一个函数，并评估输出是否与预期的一样。对于简单函数，预期结果通常是显而易见的（例如，在仅包含少量主体的示例数据上执行倾向得分匹配时）；对于更复杂的函数，可以使用 R 中其他可用函数的组合来生成预期结果（例如，通过将简单问题的结果与 R 中的其他回归事务进行比较，对我们的大型回归引擎 Cyclops 进行测试）。我们的目标是使这些测试覆盖每一行可执行源代码。

当软件包发生更改时（特别是将更改推送到软件包仓库时），将自动执行这些测试。 测试期间发现任何错误都会自动触发电子邮件发送给工作组的领导，并且必须在发布新版本的程序包之前解决错误。这些测试的源代码和预期结果的可视情况在应用程序中可以查看和使用。这些测试也可供终端用户或系统管理员使用，并且可以在其安装过程中运行，以便于在安装有关方法库时提供准确、可靠、一致的文档和客观凭证。

17.3.2 仿真

对于更复杂的函数，我们常常不是很清楚对于给定输入的预期输出是什么，在这种情况下，有时会使用仿真在给定的特定统计模型中生成输入，并确定其功能是否可以产生符合该已知模型的结果。例如，在 SelfControlledCaseSeries 程序包中，仿真用于验证该方法是否能够在仿真数据中检测出时间趋势，并合理对时间趋势进行建模。

17.4 总结

- 观察性研究应实现用自动化脚本对CDM数据到结果进行整体分析，以确保可重复性和透明性。
- 自定义研究代码应遵循最佳编程准则，包括抽象，封装，命名清晰和代码重复使用。
- 可以使用代码审查或双重编码来验证自定义研究代码。
- 方法库提供了可用于观察性研究的有效功能。
- 使用旨在开发有效软件的软件开发过程以及测试来验证方法库。

参考文献

1. Martin, Robert C. 2008. Clean Code: A Handbook of Agile Software Craftsmanship. 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR.

59. https://github.com/OHDSI/CohortMethod/issues
60. http://forums.ohdsi.org/
61. https://ohdsi.github.io/MethodsLibrary/
62. https://github.com/

第十八章　方法有效性（Method Validity）

章节负责人（Martijn Schuemie）

当我们考虑“方法有效性”的时候，其实是想回答如下的问题：

这种“方法”能否有效地解决问题？

这里所说的“方法”不仅包括了研究设计，同时也包括数据采集和整个设计的实施过程。因此，方法有效性多少有点笼统，如果没有良好的数据质量、临床有效性和软件有效性，通常就不可能观测到良好的方法有效性。在我们考虑方法有效性之前，这些证据质量方面的问题都应该提前单独考虑。

评估方法有效性的核心工作就是评估分析中的重要假设有没有被证实。例如，我们假设：倾向评分匹配（PSM, Propensity-Score Matching）可以使两个总体具有可比性，我们就需要去评估是否真是这样。在条件允许的情况下，我们通过实证检验来验证这些假设。例如，我们可以生成诊断方法来证实，在匹配之后，两个人群在各种特征上确实具有可比性。在 OHDSI 数据库里面已经有了很多进行分析时所需要生成和评估的标准化诊断方法。

在这一章中，我们将重点讨论在人群水平估计中使用的方法的有效性。我们将首先简要突出介绍一些针对于研究设计的诊断方法，然后讨论适用于大多数(如果不是所有)人群水平估计研究的诊断方法。接下来会分步描述如何使用 OHDSI 工具来执行诊断方法。本章最后，我们会讲到一部分进阶内容，对 OHDSI 方法基准的回顾及其在 OHDSI 方法库中的应用。

18.1 针对研究设计的诊断方法

对于每个研究设计，都有针对于该设计的诊断方法。这些诊断方法很多已经实现，并且可以在 OHDSI 方法库的 R 包中可以直接找到并拿来应用。例如，在 12.9 节列出了由 CohortMethod 包提供的大量诊断方法，包括：

倾向评分分布(Propensity score distribution)：用于评估不同队列的初始可比性。

倾向模型(Propensity model)：用于识别应该从模型中剔除的潜在变量。

协变量平衡(Covariate balance)：用于评估倾向评分调整是否已使队列可比(通过基线协变量测量)。

消耗(Attrition)：观察有多少受试者在不同的分析步骤中被剔除，这样可以看出结果对最初目标队列的概化。

功效(Power)：评估是否有足够的数据来回答这个问题。

Kaplan Meier 曲线(Kaplan Meier curve)：用于评估典型的发病时间，以及 Cox 模型所隐含的比例假设是否得到满足。

其他研究设计需要不同的诊断方法来测试这些设计中的各种不同的假设。例如，对于自身对照病例系列（SCCS, Self-Controlled Case Series）设计，我们需要检查观察终止与结果无关这个必要假设。这个假设常常在发生严重的、潜在致死性的事件(如心肌梗死)时被违背。我们可以通过生成类似 18.1 这样的图示来评估假设是否成立，图中分别展示了被审查的和未被审查的病例观察周期结束时间的直

方图。在我们的数据中，我们认为观察期一直延续到数据采集结束期的病例（整个数据库停止观察的日期，例如停止数据提取的日期或者研究截止的日期）是未被审查者，其他的是被审查者。在图 18.1 中我们可以看到两者的分布只有微小的差别，说明假设是成立的。

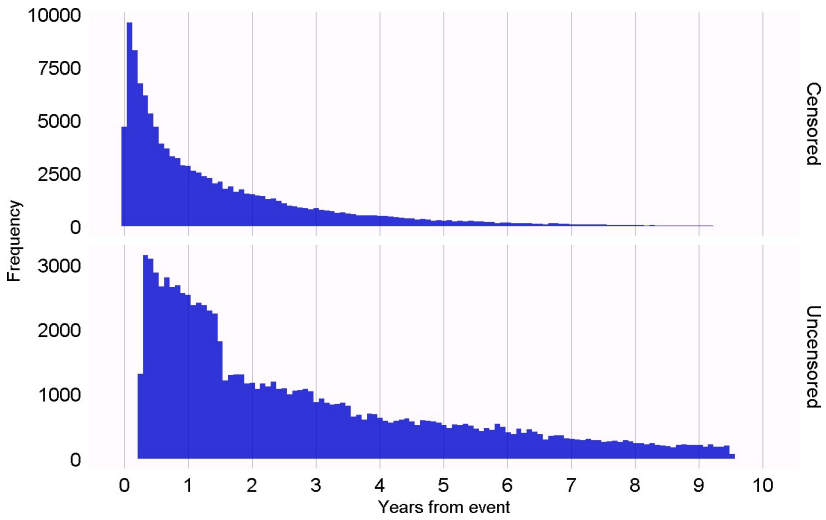


图 18.1 删失的和未删失的病例观察周期结束时间的直方图

18.2 可应用于所有估计的诊断方法

除了针对研究设计的诊断方法，还有几种诊断方法适用于所有因果效应估计方法。在所研究的问题答案已知的情况下，许多诊断方法需依靠对照假设。基于对照假设，我们可以评估研究设计产生的结果是否符合事实。对照分为阴性对照和阳性对照。

18.2.1 阴性对照

阴性对照是已知不存在因果效应的暴露-结果对，它包括阴性对照或“伪造终点”(Prasad and Jena, 2013)、选择偏倚和测量误差(Arnold et al., 2016)的手段，其中“伪造终点”被推荐作为检测混淆的一种方法(Lipsitch et al., 2010)。例如在一项关于儿童期疾病与后期发生多发性硬化的关系研究中(Zaadstra et al., 2008)，作者用了他认为不会引起多发性硬化的三种疾病做阴性对照：手臂骨折、脑震荡和扁桃体切除。结果这三种对照中有两种与多发性硬化有统计学意义上的相关性，提示研究可能存在偏倚。

我们应该选择与我们的目标假设相似的阴性对照，这意味着我们通常选择与目标假设有相同的暴露条件(所谓的“结果对照”)或相同的结果(“暴露对照”)的暴露-结果对。阴性对照需要进一步满足以下标准：

暴露不应导致结果。思考因果关系的一种方法是去“反事实”推断：与患者暴露相比，如果患者未暴露，是否会导致(或避免)结果的发生？有时这是清楚的，比如服用 ACEi 类药物会引起血管性水肿。有时候因果关系没那么显而易见，例如一种可能导致高血压的药物可以间接导致由高血压引起的心血管疾病。

暴露也不应对结果有预防或治疗作用。如果我们确信真实效应量(如风险比)为 1，这只是另一种我

们要避免的因果关系。

数据中需要有足够数量的阴性对照。我们尝试根据发生率对候选的阴性对照进行优先排序来达到这个目标。

理想的阴性对照应该是独立的，例如，我们应该避免使用有从属关系(如“指甲内生”和“脚趾甲内生”)，或是近似关系(如“左股骨骨折”和“右股骨骨折”）的对照作为阴性对照。

阴性对照最好也存在偏倚，例如一个人的社会安全号的最后一个数字基本上是一个随机数，不太可能存在混杂。因此，它不应该被用来做阴性对照。

一些人认为，阴性对照应该具有与目标暴露-结果对相同的混杂结构(Lipsitch et al., 2010)。然而，我们认为这种混杂结构是不可知的，现实中各种变量的关系往往比人们想象的复杂的多。此外，即使混杂结构已知，也不可能存在混杂结构完全相同而不存在直接因果关系的阴性对照。基于这个原因，在 OHDSI 中我们纳入大量的阴性对照，假设这样的对照组包含各种不同类型的偏倚，包括在目标假设中出现的那些偏倚。

暴露与结果之间缺乏因果关系的结论很少有文献记录。而我们通常有这样的假设：没有证据证实因果关系的存在就意味着不存在因果关系。如果对暴露和结果都进行了广泛地研究，那么这种假设成立的可能性很大，因此这种关系很可能已经被发现了。例如，一种全新的药物缺乏与某种后果有因果关系的证据可能是由于我们缺乏相关知识，而不是两者没有因果关系。基于这一原则，我们开发出一种半自动的选择阴性对照的程序(Voss et al., 2016)。简而言之，通过文献、产品标签、自发报告的信息被自动提取生成一个阴性对照候选列表。这个列表必须经过人工审查，不仅要验证自动提取是否准确，还要增加一些附加标准，如生物合理性等。

18.2.2 阳性对照

要了解当真实相对风险小于或大于 1 时方法的行为，用零值显然是不真实的，需要使用阳性对照。不幸的是，对于观察性研究来说，真实的阳性对照是有问题的，原因有三。首先，在大多数研究环境中缺乏与该特定环境相关的阳性对照，例如在比较两种治疗的效果时；第二，即使有阳性对照可用，我们对其效应值也不可能知道的十分准确，而且往往受到其测量的人群的影响；第三，当治疗方法被广泛认为会导致某种特定的结果时，会影响医生开出治疗处方时的行为，例如通过采取某些行动来降低不希望的结果出现的风险，从而使阳性对照作为评估手段变得毫无用处(Noren et al., 2014)。

因此，在 OHDSI 中，我们通过在风险暴露期间将阴性对照注入额外的模拟结果来改造成阳性对照(Schuemie et al., 2018a)。例如，假设在暴露于 ACEi 期间，我们观察到 n 次阴性对照结果“指甲内生”。如果我们现在在暴露期间再增加 n 次模拟事件，我们的风险就会增加一倍。由于这是一个阴性对照，相对于反事实的风险是 1，而注入后，它就变成了 2。

一个重要的问题是保持混杂性。阴性对照组可能表现出很强的混杂性，但如果我们随机注入额外的结果，这些新的结果将不会具有混杂性，因此我们可能会过于乐观地估计我们处理阳性对照组混杂性的能力。为了保持混杂性，我们希望新的结果与原始结果显示基线对象特异性协变量相似的关联。为了达到这个目的，我们对每个结果都训练了一个模型，用暴露前获得的协变量来预测在暴露期间与结果相关的存活率。这些协变量包括人口学特征，以及所有的诊断、药物暴露、测量结果和医疗程序记录。利用 10 折交叉验证来选择正则化超参数的 L1 正则化 Poisson 回归适用于这个预测模型(Suchard et al., 2013)。然后，我们在暴露期间使用预测的概率抽样模拟结果，将真实效应量增加到所需要的幅度。因

此，最终的阳性对照同时包含了真实的和模拟的结果。

图 18.2 描述了这个过程。请注意，虽然这个过程模拟了几个重要的偏倚来源，但它并不能获得所有的偏倚。例如，一些测量误差的影响是不包含在内的。合成的阳性对照的阳性预测值和敏感性为常数，但这在实际中不一定成立。

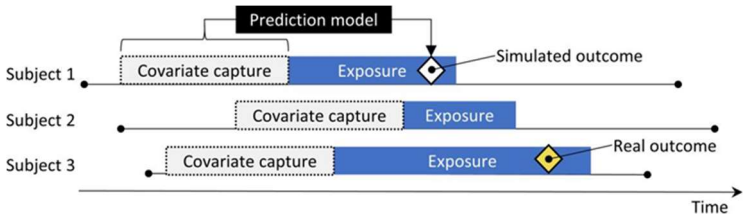


图 18.2：从阴性对照合成阳性对照。

虽然我们对每个对照使用了同样的真实“效应量”，但是不同处理效果的统计数据是用不同的方法估计的。我们认为阴性对照不存在因果效应，那么所有统计数据，包括相对风险、风险比、优势比、发生率比，不论是条件的还是边际的，以及处理组的平均处理效果(ATT)和总体平均处理效果(ATE)都为 1。在创建阳性对照的过程中，我们用一个以患者中保持恒定的比率为条件的模型，采用了在不同时间和患者之间保持恒定的发生率来合成结果，直到实现边际效应。因此，真实效应量被保持为处理组的边际发生率。假设我们在合成结果的过程中使用的模型是正确的，那么条件效应量和 ATE 也是恒定的。由于所有结果都是罕见的，优势比与相对风险几乎完全相同。

18.2.3 实证评估

根据针对阴性和阳性对照的特定方法的评估，我们可以通过计算一系列指标来了解操作特征，例如：

- **受试者工作曲线下的面积（AUC）**：区分阳性对照和阴性对照的能力。
- **覆盖率**：真实效果量在 95%置信区间内的频率。
- **精度均值**：精度用 1/(标准误差)2 来计算，越高的精度表示越窄的置信区间。我们使用几何平均值使得精度呈偏态分布。
- **均值平方误差（MSE）**：效果量点估计的对数与真实效果量的对数之间的均值平方误差。
- **1 类错误**：在阴性对照中，零假设被拒绝的概率(当α= 0.05 时)。这相当于假阳性率，等于 1-特异性。
- **2 类错误**：在阳性对照中，零假设不被拒绝的概率(当α= 0.05 时)。这相当于假阴性率，等于 1-敏感性。
- **不可估计**：方法不能估计的对照有多少？可能有许多种原因导致无法进行估计，例如，倾向评分匹配后没有剩余对象，或剩余对象都没有结果出现。

根据我们的用例，我们可以评估这些操作特性是否适合我们的目标。 例如，如果我们希望执行信号检测，我们可能会关心 1 类错误和 2 类错误，或者如果我们想调整α阈值，可能会检查 AUC。

18.2.4 P 值校准

1 类错误（在α=0.05 时）通常大于 5%。换句话说，当零假设成立时，我们通常有超过 5%的可能性拒绝它。原因就是 p 值仅反映随机误差（该误差是由样本数量有限造成的），而不反映系统误差（例如由于混淆因素引起的误差）。OHDSI 开发了一种校准 p 值的方法，以将 1 类错误恢复为标称值。（Schuemie et al.，2014）我们从阴性对照的实际效应估计中得出经验零分布。这些阴性对照的估计为我们展示了当零假设成立时的情况，并且它们可以被用来估计经验零分布。

形式上，我们用估计拟合高斯概率分布，同时考虑每个估计的采样误差。令θˆi 表示来自第 i 个阴性对照药物-结果对的效应量估计值的对数值（相对风险，优势比或发生率比），同时令τˆi 表示相应的标准差估计值，其中 i = 1, . . . , n。用θi 表示真实效应量的对数值(假设阴性对照为 0)，βi 表示与第 i 个对相关的真实偏倚(未知)，即，真实效应量的对数值与研究给出的对照 i 估计值的对数值的差别可以无穷大。与标准 p 值计算一样，我们假设θˆi 服从均值为θi + βi 且标准差为τˆ2 的正态分布。请注意，在传统的 p 值计算中βi 总是假设等于零，但这里我们假设βi 服从均值 μ 和方差σ2 的正态分布。这就代表了零（偏差）分布。我们可以通过最大似然估计得到 μ 和σ2。总而言之，我们假设：

$$\beta_i \sim N(\mu, \sigma^2) \text{ and } \hat{\theta}_i \sim N(\theta_i + \beta_i, \tau_i^2)$$

其中 N (a, b) 表示均值为 a 方差为 b 的一个高斯分布，并且可以通过最大化以下的似然函数来估计 μ 和σ2：

$$L(\mu, \sigma|\theta, \tau) \propto \prod_{i=1}^{n} \int p(\hat{\theta}_i|\beta_i, \theta_i, \hat{\tau}_i) p(\beta_i|\mu, \sigma) \mathrm{d}\beta_i$$

最终得到最大似然估计 μˆ 和 σˆ。接下来我们用经验零分布来计算校准的 p 值。令 θˆi+1 表示从一个新的药物-结果对得到的效应量估计值的对数值，同时令τˆn+1 表示相应的标准差估计值。根据上述假设并假设βn+1 来自相同的零分布，我们得到以下公式：

$$\hat{\theta}_{n+1} \sim N(\hat{\mu}, \hat{\sigma} + \hat{\tau}_{n+1})$$

当θˆn+1 小于 μˆ时，新对的单侧校准 p 值为：

$$\phi\left(\frac{\theta_{n+1} - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{n+1}^2}}\right)$$

其中 ϕ(·)表示标准正态分布的累积分布函数。当θˆn+1 大于 μˆ时，单侧校准 p 值为:

$$1 - \phi\left(\frac{\theta_{n+1} - \hat{\mu}}{\sqrt{\hat{\sigma}^2 + \hat{\tau}_{n+1}^2}}\right)$$

18.2.5 置信区间校准

同样地，我们通常能观察到 95%置信区间的覆盖范围小于 95%，真正的效应量在 95%置信区间的概率少于 95%。对于置信区间校准（Schuemie et al., 2018a），我们还通过使用阳性对照扩展了 p 值校准的框架。通常，校准后的置信区间比标称置信区间更宽，这反映了标准程序中未考虑但在校准中考虑到的问题(如未测量的混杂、选择偏倚和测量误差)，但这也不是一定的。

形式上，我们假设与第 i 个对所对应的偏倚βi，也服从高斯分布，但这次的均值和标准差与θi 线性相关，真实的效应量为：

βi ~ N (μ(θi), σ2(θi))

当 μ(θi) = a + b × θi 并且 σ(θi)2 = c + d× | θi |时，

我们使边际似然函数最大化来估计 a, b, c 和 d，在其中将未观察到的βi 积分出来：

$$l(a,b,c,d|\theta,\hat{\theta},\hat{\tau}) \propto \prod_{i=1}^{n} \int p(\hat{\theta}_i|\beta_i,\theta_i,\hat{\tau}_i)p(\beta_i|a,b,c,d,\theta_i)\mathrm{d}\beta_i$$

得到最大似然估计(aˆ, ˆb, cˆ, dˆ)。

我们利用系统误差模型计算得到校准后的置信区间。将一个新的目标结果的效应量估计值的对数值设为θˆn+1，并用τˆn+1 表示对应的估计标准误。根据以上假设，并假设βn+1 来自相同的系统误差模型，我们有：

$$\hat{\theta}_{n+1} \sim N(\theta_{n+1} + \hat{a} + \hat{b} \times \theta_{n+1}, \hat{c} + \hat{d} \times \mid \theta_{n+1} \mid + \hat{\tau}_{n+1}^2)$$

通过求解θn+1 的这个方程，我们找到了经过校准的 95%CI 的下限：

$$\Phi\left(\frac{\theta_{n+1} + \hat{a} + \hat{b} \times \theta_{n+1} - \hat{\theta}_{n+1}}{\sqrt{(\hat{c} + \hat{d} \times \mid \theta_{n+1} \mid) + \hat{\tau}_{n+1}^2}}\right) = 0.025,$$

其中Φ(·)表示标准正态分布的累积分布函数。类似地，我们用相似的方法在概率为 0.975 时求出上界。并用概率为 0.5 时定义校准后的点估计。

p 值校准和置信区间的校准都可在 EmpiricalCalibration 软件包中实现。

18.2.6 跨站点复制

方法验证的另一种形式是在多种不同的数据库上展开研究，这些数据库代表了不同人群、不同卫生保健系统，和/或不同数据采集过程。先前的研究表明，在不同的数据库中执行相同的研究设计会产生完全不同的效应量估计值(Madigan et al., 2013b)。这意味着要么是不同人群中的效应量差异很大，要么是研究设计没有充分考虑在不同数据库中存在的不同偏倚。事实上，我们发现，通过对置信区间的经验校准来处理数据库中的残差偏倚可以极大地减少研究之间的异质性（Schuemie et al., 2018a)。

I2 评分是表示数据库间异质性的一种方法，它描述了由异质性而非偶然性导致的变异在研究之间的总变异中的百分比(Higgins et al., 2003)。 尽管可以用 25%，50%和 75%的 I2 值来表示轻度、中度和高度异质性，但对 I2 值的简单分类并不适用于所有情况。在一项评估抑郁症治疗效果的研究中，研究者使用了大规模倾向评分调整的新用户队列设计（Schuemie et al., 2018b)，仅观察到 58%的 I2 值低于 25%。经过经验性校准后，这一比例提高到了 83%。

人观察数据库之间的异质性使人们对估计的有效性产生怀疑遗憾的是，这种反推并不正确。不对异质性进行观察并不能保证你得到一个无偏估计。如果所有的数据库都有相同的偏倚，那么所有估计都存在一样的错误，那是不可能的。

18.2.7 灵敏度分析

当设计研究时，通常会有不确定的设计选择。例如，是否应使用分层的倾向评分匹配？如果使用分层，需要分多少层？什么是合适的风险暴露时间？面对这种不确定性时，一种解决方案是评估各种选择，并观察结果对设计选择的敏感性。如果各种选择下的估计值保持不变，那么我们可以说该研究具有不确定性。

对敏感性分析的定义不应与其他人使用的定义相混淆，例如 Rosenbaum（2005），后者将敏感性分析定义为“评估研究的结论如何被不同程度的隐藏偏倚所改变”。

18.3 实践中的方法验证

这里我们以第 12 章中的示例为基础，在该示例中，我们以噻嗪和噻嗪类利尿剂（THZ）作为对比，研究了 ACE 抑制剂（ACEi）的使用对血管性水肿发生风险和急性心肌梗塞（AMI）的影响。在第 12 章中，我们已经探讨了许多特定针对我们所使用的研究设计的诊断方法，即队列方法。在这里，我们还可以应用一些其他诊断方法，如果使用其他设计，这些诊断方法也可以应用。如果按照第 12.7 节中的描述使用 ATLAS 实施研究，则可在 ATLAS 生成的研究 R 包中包含的 Shiny 应用程序里使用这些诊断方法。如果研究是用 R 实现的（如第 12.8 节所述），那么就可以用不同包中的可用 R 函数，这会在下一节中讲到。

18.3.1 选择阴性对照

我们需要选择阴性对照，即认为不存在因果关系的暴露-结果对。对于我们的示例研究这类的效果比较估计，我们要选择那些既不是由目标物引起的，也不是由对比者暴露引起的阴性对照结果。我们希望能够有充足的阴性对照，以确保我们的对照中包含了各种偏倚，同时也允许进行经验性校准。根据经验，我们通常需要有 50-100 个这样的阴性对照。我们一般可以完全手动地设定这些对照，但幸运的是，ATLAS 提供了相关功能，让我们可以使用文献、产品标签和自发报告中的数据来辅助选择阴性对照。

要生成阴性对照的候选列表，我们首先需要创建一个包含所有相关的暴露因素的概念集。在本例中，我们选择了 ACEi 和 THZ 类中的所有成分，如图 18.3 所示。

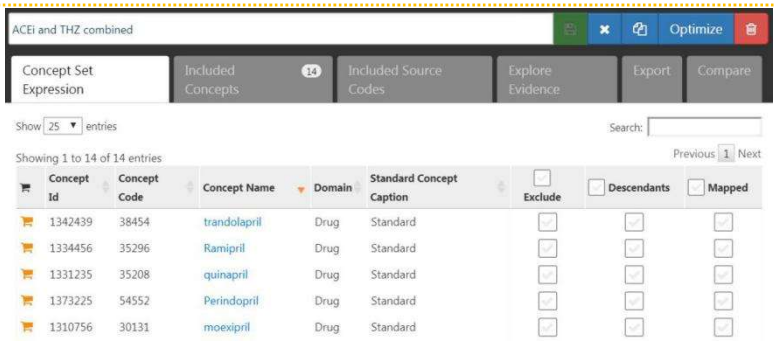


图 18.3: 一个包含目标概念和对比者暴露概念的概念集。

接下来，我们进入“Explore Evidence”选项卡，然后点击 Generate 按钮。生成证据概述需要几分钟，之后您可以点击 View Evidence 按钮。它将展示结果列表，如图 18.4 所示。

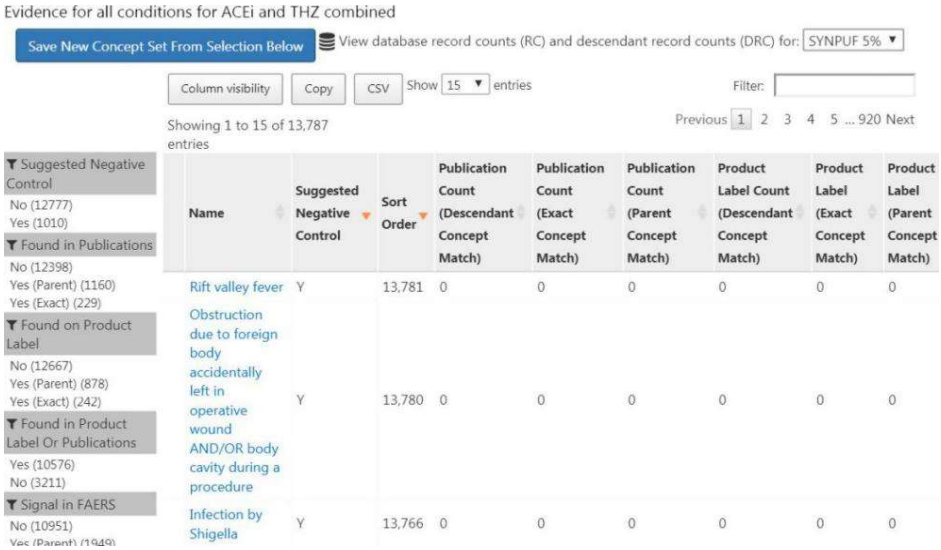


图 18.4: 候选对照结果，包含文献，产品标签和自发报告中发现的证据的概述。

此列表显示了各种状况的概念，以及将这种状况与我们定义的所有暴露相关联的证据的概述。例如，我们可以看到通过各种策略在 PubMed 中找到的所有将暴露与结果关联起来的文献数量，把我们的目标暴露列为不良反应的产品标签以及自发报告的数量。默认情况下，列表将首先显示候选阴性对照。然后可选择“Sort Order”对其进行排序，其主要表现该状况在观察数据库集合中的发生率。排序顺序越高，发生率越高。虽然在这些数据库中的发生率可能与我们希望进行研究的数据库中的发生率不一致，但它可能是一个很好的近似值。

下一步是手动检查候选列表，通常从顶部开始，也是从发生率最高的状况开始，然后逐步向下查看，直到检查满足我们的要求为止。一种典型的实现方法是将列表导出到 CSV（逗号分隔值）文件中，并让临床医生根据 18.2.1 节中提到的标准对其进行审查。

对于我们的示例研究，我们选择了在附录 C.1 中列出的 76 个阴性对照。

18.3.2 添加对照

一旦我们确定了一组阴性对照，就必须将其纳入到我们的研究当中。我们必须定义一些逻辑来将阴性对照状况概念集转化为结果队列。12.7.3 节讨论了 ATLAS 如何根据用户的几个必要选择创建出这样的队列。通常，我们只要根据阴性对照概念或其子类的发生情况来创建队列。如果该研究是由 R 实现的，那么可以使用 SQL（结构化查询语言）来构建阴性对照队列。第 9 章讲述了如何通过 SQL 和 R 来创建队列。我们就将如何编写适用的 SQL 和 R 作为练习留给读者了。

OHDSI 工具还提供了自动从阴性对照生成阳性对照并放入队列的功能。这个功能可以在第 12.7.3 节描述的 ATLAS 中的评估设置部分中被找到。该功能由 synthesizePositiveControls 函数实现，这个函数在 MethodEvaluation 包中。在这里我们使用生存模型为每个阴性对照生成了对应的 3 个阳性对照，真实效应量分别为 1.5，2 和 4：

```
library(MethodEvaluation)
# Create a data frame with all negative control exposure-
# outcome pairs, using only the target exposure (ACEi = 1).
eoPairs <- data.frame(exposureId = 1,
                      outcomeId = ncs)

pcs <- synthesizePositiveControls(
  connectionDetails = connectionDetails,
  cdmDatabaseSchema = cdmDbSchema,
  exposureDatabaseSchema = cohortDbSchema,
  exposureTable = cohortTable,
  outcomeDatabaseSchema = cohortDbSchema,
  outcomeTable = cohortTable,
  outputDatabaseSchema = cohortDbSchema,
  outputTable = cohortTable,
  createOutputTable = FALSE,
  modelType = "survival",
  firstExposureOnly = TRUE,
  firstOutcomeOnly = TRUE,
  removePeopleWithPriorOutcomes = TRUE,
  washoutPeriod = 365,
  riskWindowStart = 1,
  riskWindowEnd = 0,
  endAnchor = "cohort end",
  exposureOutcomePairs = eoPairs,
  effectSizes = c(1.5, 2, 4),
  cdmVersion = cdmVersion,
  workFolder = file.path(outputFolder, "pcSynthesis"))
```

```
  firstOutcomeOnly = TRUE,
  removePeopleWithPriorOutcomes = TRUE,
  washoutPeriod = 365,
  riskWindowStart = 1,
  riskWindowEnd = 0,
  endAnchor = "cohort end",
  exposureOutcomePairs = eoPairs,
  effectSizes = c(1.5, 2, 4),
  cdmVersion = cdmVersion,
  workFolder = file.path(outputFolder, "pcSynthesis"))
```

请注意，我们必须模拟在评估研究设计中使用的风险暴露时间设置。synthesizePositiveControls 函数将提取有关暴露和阴性对照结果的信息，拟合每个暴露-结果对的结果模型，并合成结果。阳性对

照结果队列会被添加到 cohortDbSchema 和 cohortTable 指定的队列表中。最终得到的 pcs 数据框架包含了关于合成阳性对照的信息。

接下来，我们必须执行相同的研究，来估计目标变量的效应，并同时估计阴性对照和阳性对照的效应。在 ATLAS 的比较对话框中设置一组阴性对照，用来指导 ATLAS 计算这些对照的估计值。类似地，在评估设置中生成的阳性对照也加入到我们的分析中。在 R 里，阴性对照和阳性对照应与其他结果同样处理。OHDSI 方法库中的所有估计功能包都可以有效地估计许多效应。

18.3.3 性能验证

图 18.5 展示了示例研究中对阴性和阳性对照的估计效应量，并按真实效应量分层。这些散点图包含在由 ATLAS 生成的研究 R 包附带的 Shiny 应用程序中，可以使用 MethodEvaluation 包的 plotControls 函数生成。请注意，对照的数量通常少于所定义的数量，因为没有足够的数据来兼顾生成估计值和合成阳性对照。

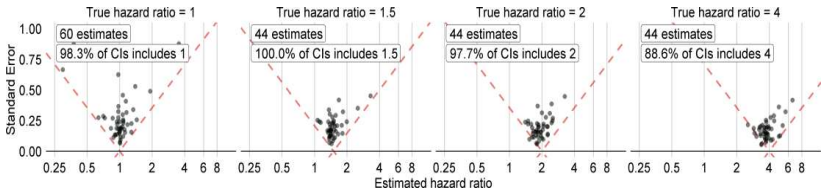


图 18.5：阴性对照(真实风险比= 1)和阳性对照(真实风险比> 1)的估计值。

每个点表示一个对照。虚线以下的点表示真实效应量不在估计的置信区间内。

基于这些估计，我们使用 MethodEvaluation 包中的 computeMetrics 函数计算表 18.1 所示的指标。

表 18.1：由阴性和阳性对照估计中得出的方法性能指标

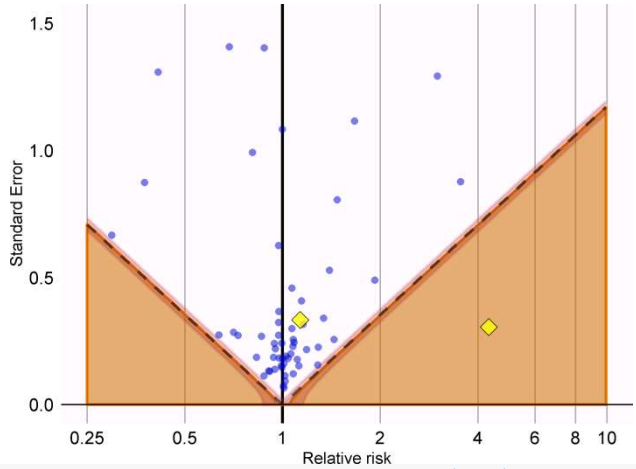
Metric	Value
AUC	0.96
Coverage	0.97
Mean Precision	19.33
MSE	2.08
Type 1 error	0.00
Type 2 error	0.18
Non-estimable	0.08

我们看到覆盖率（Coverage）和 1 类错误（Type 1 error）非常接近其 95%和 5%的标称值，而且 AUC 非常高。当然，情况并非总是如此。

注意，尽管在图 18.5 中，当真实危险比为 1 时，并非所有置信区间都包括 1，但表 18.1 中的 1 类错误（Type 1 error）为 0%。这是一种例外情况，原因是 Cyclops 包中的置信区间是似然分析法估计的，这种方法比传统方法更精确，但会导致置信区间不对称。而 p 值是在假设置信区间对称的情况下计算的，并被用于计算 1 类错误。

18.3.4 P 值校准

我们可以用阴性对照的估计值来校准 p 值。这是在 Shiny 应用程序中自动完成的，也可以用 R 手动完成。假设我们已经创建了第 12.8.6 节中描述的摘要对象 summ，我们可以绘制经验性校准效果图：




```
# Estimates for negative controls (ncs) and outcomes of interest
(ois):
ncEstimates <- summ[summ$outcomeId %in% ncs, ]
oiEstimates <- summ[summ$outcomeId %in% ois, ]

library(EmpiricalCalibration)
plotCalibrationEffect(logRrNegatives = ncEstimates$logRr,
                      seLogRrNegatives = ncEstimates$seLogRr,
                      logRrPositives = oiEstimates$logRr,
                      seLogRrPositives = oiEstimates$seLogRr,
                      showCis = TRUE)
```

图 18.6：P 值校准：虚线以下的估计值表示未校准的 p<0.05。阴影区域的估计值表示校准后的 p<0.05。阴影区域边缘的窄带表示 95%可信区间。点表示阴性对照。菱形表示目标结果。

在 18.6 中，我们看到阴影区域与虚线所标出的区域几乎完全重叠，这表明阴性对照组几乎没有观察到任何偏倚。其中一个目标结果（急性心肌梗死）在虚线和阴影区域上方，这表明不管是根据未校准的 P 值还是校准后的 P 值，我们都不能拒绝零假设。另一个结果（血管性水肿）在阴性对照中明显突出，并且很好地落在未校准和校准 p 值均小于 0.05 的区域内。

我们可以计算校准后的 p 值：

```
null <- fitNull(logRr =ncEstimates$logRr,
                seLogRr = ncEstimates$seLogRr)
calibrateP(null,
           logRr=
           oiEstimates$logRr,
           seLogRr           =
           oiEstimates$seLogRr
           )
```

```
## [1] 1.604351e-06 7.159506e-01
```

并与未校准的 p 值进行对比：

```
oiEstimates$p
```

[1] [1] 1.483652e-06 7.052822e-01

正如预期的那样，因为几乎没有观察到偏倚，未校准和校准后的 p 值非常接近。

18.3.5 置信区间校准

类似地，我们可以使用阴性和阳性对照的估计值来校正置信区间。Shiny 程序可以自动报告校准后的置信区间。在 R 中，我们可以使用 EmpiricalCalibration 包中的 fitSystematicErrorModel 和 calibrateConfidenceInterval 函数来校准区间，这在“置信区间的经验性校准”简介中有详细描述。

在校准前，血管性水肿和急性心肌梗死的风险比估计值 (95%置信区间) 分别为 4.32 (2.45-8.08) 和 1.13（0.59-2.18），而校准后的风险比则分别是 4.75（2.52-9.04）和 1.15（0.58-2.30）。

18.3.6 数据库之间异质性

正如我们在一个数据库上执行分析一样，在本例中我们用的是 IBM 的 MarketScan Medicaid (MDCD)数据库，我们也可以在符合通用数据模型（CDM）的其他数据库上运行相同的分析代码。图 18.7 显示了在五个数据库之上就血管性水肿进行的分析所得的森林图和荟萃分析估计值（假设为随机效应）（DerSimonian 和 Laird，1986）。该结果是由使用 EvidenceSynthesis 包中的 plotMetaAnalysisForest 函数所生成的。

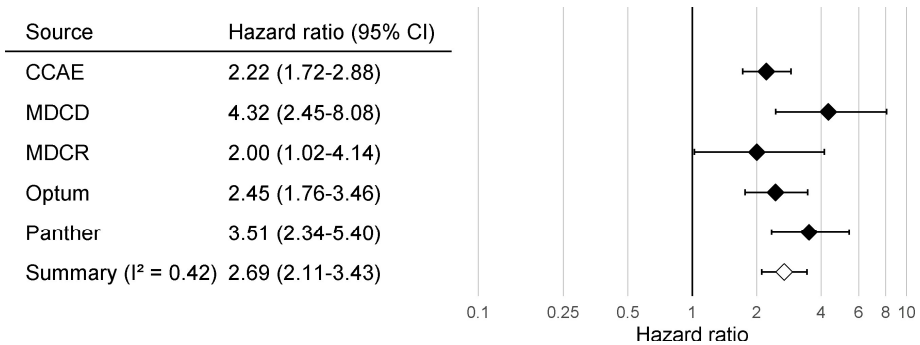


图 18.7：在比较 ACE 抑制剂与噻嗪类利尿剂对血管性水肿的风险时，我们整理了来自五个不同数据库及其荟萃分析结果的效应量估计值和 95%置信区间(CI)。尽管所有的置信区间都大于 1，表明一致认定是存在一种效应的，但 I2 值表明数据库之间存在异质性。但是如果使用校准后的置信区间来计算 I2 值（如图 18.8 所示），这种异质性可以通过对每个数据库中的阴性和阳性对照测量到的偏倚来解释。经验性校准似乎恰当地考虑了这种偏倚。

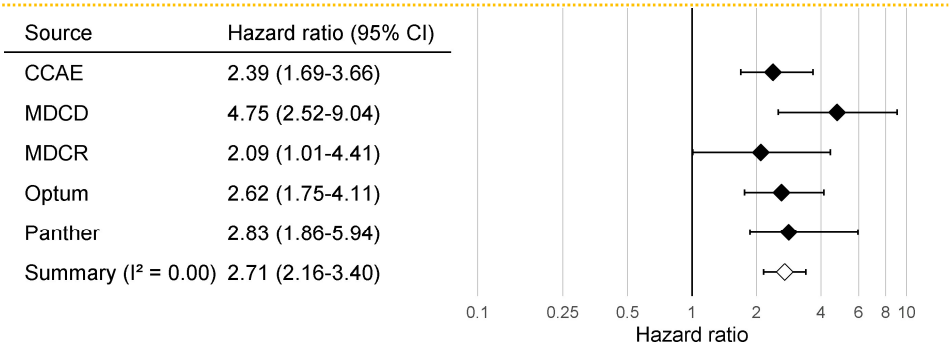


图 18.8: 来自五个不同数据库及其荟萃分析结果的效应量估计值和 95%置信区间(CI)经过校准后，用于比较使用 ACE 抑制剂与噻嗪类利尿剂对血管性水肿的风险比。

18.3.7 敏感性分析

我们分析中的一种设计选择是对倾向评分使用可变比率匹配（Matching）。 但是，我们也可以对倾向评分使用分层（Stratification）。 因为我们不确定该选择，所以我们可能决定同时使用两者。 表 18.2 显示了使用可变比率匹配和分层（有 10 个均等大小的分层）时，未校准与校准后的心肌梗死和血管性水肿的效果量的估计值。

表 18.2: 两种分析变量的未校准和校准后风险比（95%置信区间）

结果	调整方式	未校准	已校准
血管性水肿	匹配	4.32 (2.45 - 8.08)	4.75 (2.52 - 9.04)
血管性水肿	分层	4.57 (3.00 - 7.19)	4.52 (2.85 - 7.19)
急性心肌梗死（AMI）	匹配	1.13 (0.59 - 2.18)	1.15 (0.58 - 2.30)
急性心肌梗死（AMI）	分层	1.43 (1.02 - 2.06)	1.45 (1.03 - 2.06)

我们看到匹配分析和分层分析得出的估计值非常一致，分层的置信区间完全落在匹配的置信区间内。这表明我们对设计选择的不确定性不会影响估计的有效性。分层似乎确实给了我们更强的效力（更窄的置信区间），这并不奇怪，因为匹配会导致数据丢失，而分层却不会。由于层内残余的混杂，这样做的代价可能是偏倚的增加，尽管我们没有在校准的置信区间中看出能反映偏倚增加的证据。

研究诊断方法使我们甚至可以在完全执行研究之前评估设计选择。建议不要在生成和检查所有研究诊断之前确定最终方案。为避免 p 值操纵（p-hacking）（调整设计以获得期望的结果），应在目标变量的效果

18.4 OHDSI 方法基准

尽管推荐的做法是在实际应用环境中评估方法的性能，在本研究所使用的数据库中选择那些与目

标暴露-结果对在某些方面相似的阴性和阳性对照（例如使用相同的暴露条件或是结果相同），而评价一种方法在通用条件下的性能也是有价值的。这就是我们建立 OHDSI 方法评估基准的原因。基准测试使用一系列对照问题来评估性能，这些对照问题包括那些具有慢性或急性结果以及长期或短期暴露的问题。该基准测试的结果有助于证明方法的总体效用，并且当还没有针对特定环境下的实证评价时，可以用来形成对方法性能的先验信心。这个基准包括 200 个经过精心挑选的阴性对照，这些阴性对照可以分为八个类别，每个类别中的对照共享相同的暴露或相同的结果。如第 18.2.2 节所述，从这 200 个阴性对照中合成 600 个阳性对照。为了评估一种方法，必须使用该方法为所有对照产生效应量估计值，然后才能计算第 18.2.3 节中描述的指标。该基准是公开可用的，并且可以按照 MethodEvaluation 包的 Running the OHDSI Methods Benchmark 简介中的描述进行部署。

我们已经通过这个基准测试运行了 OHDSI 方法库中的所有方法，每个方法都有不同的分析选择。例如，使用倾向评分匹配、分层和加权来评估队列方法。该实验是在四个大型观察性医疗卫生数据库上执行的。结果可于在线 Shiny 程序 1 中查看，结果表明，尽管一些方法显示出较高的 AUC（区分阳性对照和阴性对照的能力），但是大多数情况下，这些方法都显示了较高的 1 类错误和 95%置信区间的低覆盖率，如图 18.9 所示。

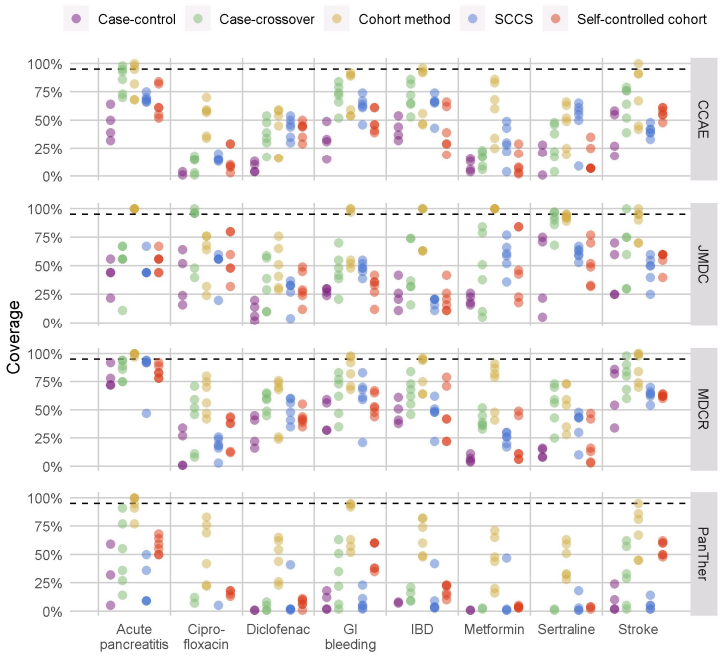


图 18.9：方法库中方法的 95%置信区间的范围。

每个点代表一组特定分析选择的性能。

虚线表示标称性能（95%覆盖率）。

SCCS =自控病例系列，GI =胃肠道，IBD =炎症性肠病。

18.5 总结

– 方法的有效性取决于是否满足该方法的基本假设。
– 在可能的情况下，应使用研究诊断方法对这些假设进行实证检验。
– 对照假设，即结果已知的问题，应当用来评估一个特定的研究设计产生的结果是否符合事实。
– 通常，p 值和置信区间并不表现出对照假设测量得到的标称特征。
– 通常可以使用经验性校准将这些特征恢复为标称值。
– 研究诊断可用于指导分析设计选择和方案调整，这需要研究人员保持对目标效应不可见，以避免 P 值篡改的发生。

参考文献

1. Arnold, B. F., A. Ercumen, J. Benjamin-Chung, and J. M. Colford. 2016. “Brief Report: Negative Controls to Detect Selection Bias and Measurement Bias in Epidemiologic Studies.” Epidemiology 27 (5): 637–41.
2. DerSimonian, R., and N. Laird. 1986. “Meta-analysis in clinical trials.” Control Clin Trials 7 (3): 177–88.
3. Higgins, J. P., S. G. Thompson, J. J. Deeks, and D. G. Altman. 2003. “Measuring inconsistency in meta-analyses.” BMJ 327 (7414): 557–60.
4. Lipsitch, M., E. Tchetgen Tchetgen, and T. Cohen. 2010. “Negative controls: a tool for detecting confounding and bias in observational studies.” Epidemiology 21 (3): 383–88.
5. Madigan, D., P. B. Ryan, M. Schuemie, P. E. Stang, J. M. Overhage, A. G. Hartzema, M. A. Suchard, W. DuMouchel, and J. A. Berlin. 2013. “Evaluating the impact of database heterogeneity on observational study results.” Am. J. Epidemiol. 178 (4): 645–51.
6. Noren, G. N., O. Caster, K. Juhlin, and M. Lindquist. 2014. “Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance.” Drug Saf 37 (9): 655–59.
7. Prasad, V., and A. B. Jena. 2013. “Prespecified falsification end points: can they validate true observational associations?” JAMA 309 (3): 241–42.
8. Rosenbaum, P. 2005. “Sensitivity Analysis in Observational Studies.” In Encyclopedia of Statistics in Behavioral Science. American Cancer Society. https://doi.org/10.1002/0470013192.bsa606.
9. Schuemie, M. 2018. “Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data.” Proc. Natl. Acad. Sci. U.S.A. 115 (11): 2571–7.

10. Schuemie, M. J., P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan. 2014. "Interpreting observational studies: why empirical calibration is needed to correct p-values." Stat Med 33 (2): 209–18.
11. Schuemie, M. J., P. B. Ryan, G. Hripcsak, D. Madigan, and M. A. Suchard. 2018. "Improving reproducibility by using high-throughput observational studies with empirical calibration." Philos Trans A Math Phys Eng Sci 376 (2128).
12. Suchard, M. A., S. E. Simpson, Ivan Zorych, P. B. Ryan, and David Madigan. 2013. "Massive Parallelization of Serial Inference Algorithms for a Complex Generalized Linear Model." ACM Trans. Model. Comput. Simul. 23 (1): 10:1–10:17. https://doi.org/10.1145/2414416.2414791.
13. Voss, E. A., R. D. Boyce, P. B. Ryan, J. van der Lei, P. R. Rijnbeek, and M. J. Schuemie. 2016. "Accuracy of an Automated Knowledge Base for Identifying Drug Adverse Reactions." J Biomed Inform, December.
14. Zaadstra, B. M., A. M. Chorus, S. van Buuren, H. Kalsbeek, and J. M. van Noort. 2008. "Selective association of multiple sclerosis with infectious mononucleosis." Mult. Scler. 14 (3): 307–13.

65. https://date. OHDSI.org/MethodEvalViewer/

第 19 章 研究步骤

章节编辑: Sara Dempster & Martijn Schuemie

本章旨在为使用 OHDSI 工具进行观察性研究的设计与实施提供一种通用的逐步指南。我们将对研究流程中的每一个步骤进行分解并进行大体的描述，在某些情况下会对一些主要研究类型的特定方面进行描述，包括：(1) 患者特征刻画；(2) 群体水平评估 (PLE)；(3) 本书前面章节中所描述的患者水平预测 (PLP)。因此，我们将在这里整合之前章节中所讨论过的内容，以便初学者理解。同时，本章也可以作为独立的章节供寻求实用的高层次解释的读者使用，为其在其他章节中寻找更深入的材料提供帮助。最后，我们将通过几个关键示例来进行说明。

此外，我们将总结 OHDSI 社区推荐的观察性研究指南和最佳实践。其中一些通用的原则与许多其他观察性研究指南中的最佳实践建议是共通的，另一些建议流程则更多的针对 OHDSI 框架。因此我们会重点介绍那些 OHDSI 工具栈所支持的特定的 OHDSI 方法。

在本章中，我们假定读者熟悉 OHDSI 工具，并且可以使用 R 和 SQL，因此本章不讨论如何设置以上基础环境 (相关指导请参见第 8 章和第 9 章)。我们假定读者想在自己的站点上，基于已经设置为 OMOP CDM 的数据库开展研究 (有关 OMOP ETL 的内容请参见第 6 章)。值得强调的是，下面讨论的研究套件准备完毕后，原则上也可以分发到其他站点并运行。针对 OHDSI 协作网研究的其他具体注意事项，包括组织和技术细节等，将在第 20 章中进行详细讨论。

19.1 通用最佳实践指南

19.1.1 观察性研究定义

观察性研究，其定义是指在特定患者治疗过程中仅观察患者，不进行干预的研究。有时观察数据的收集具有特定的目的，例如在开展临床注册研究时。但在多数情况下，这些数据的收集并不是为了解决手头上特定的研究问题，而是具有其他的目的，常见于电子病历 (EHR) 或医保索赔数据的收集。观察性研究通常被称为数据的二次利用。进行观察性研究的基本指导原则是明确研究问题，并在开展研究之前确定研究方法。在这方面，观察性研究与临床试验基本没有区别，但临床试验招募并及时跟踪患者的主要目的是回答某个特定的问题，该问题通常与治疗干预的有效性和 (或) 安全性相关。观察研究的分析方法与临床试验有很大的不同，最突出的是群体水平评估 (PLE) 的观察性研究缺乏随机性，如果研究目的是进行因果推断，则需要采取措施以控制混杂因素 (有关 OHDSI 支持的群体水平评估 (PLE) 研究设计与方法，例如通过在许多特征上平衡人群来消除观察中存在的混杂因素等方法，详见第 12 章与第 18 章)。

19.1.2 研究设计的预先规范

观察性研究设计及其参数设定的预先规范，对于避免因潜意识或有意改进方法以获得期望的结果而进一步产生偏倚 (有时也被称为 P 值篡改) 至关重要。第二次使用数据比第一次使用更倾向于不提前详细说明研究细节，因为这些数据 (例如 EHR 数据和医保索赔数据) 有时会给研究人员一种“无限可能”的感觉，从而导致研究偏倚。尽管现有数据很容易获取，但仍然要实施严格的科学研究框架。在 PLE 或 PLP 中，预先指定的原则尤其重要，以确保得到严格或可重复的结果，因为这些结果可能最终

会为临床实践或监管决策提供依据。即使纯粹出于探索原因而进行的特征刻画研究，最好也制定一个严谨明确的计划。否则，不断发展的研究设计和分析过程将难以被记录、解释和重复。

19.1.3 研究方案

观察性研究的研究计划应该存档于一个在研究实施前就创立的研究方案中。研究方案至少包括主要的研究问题、研究方法和用于回答研究问题的度量标准。研究人群的描述应细化以使其他人可以完全复制。此外，所有方法或统计过程，以及预期研究结果的形式（如矩阵、表格和图表）都应加以说明。通常，研究方案还会描述一组预分析，用于评估研究的可行性或统计效力。此外，研究方案可能包含对主要研究问题变化的说明，即敏感性分析。敏感性分析是用来评估研究设计的选择对整体研究结果的潜在影响，应尽可能提前说明。有时，在一些无法预料的情况下，需要对已经完成的研究方案进行修正。如果这种修正十分必要，那么记录方案本身被修正的内容和修正的原因至关重要。特别是在 PLE 和 PLP 研究中，完整的研究方案应该存储于一个独立的平台中（比如 clinicaltrials.gov 或 OHDSI’s study Protocols sandbox），其方案版本与任何内容更改都可以通过时间戳进行独立溯源。通常，您所在的机构或数据源的所有者会要求在实施研究之前有机会审查和批准您的方案。

19.1.4 标准化分析

OHDSI 的独特优势是其工具可以通过识别观察性研究中反复提到的几类主要问题来支持制定研究计划、存档和报告（第 2、7、11、12、13 章），从而通过对重复的部分进行自动处理来简化研究方案策划和研究实施等过程。许多工具都可以将一些研究设计或度量标准进行参数化，以应对大多数未来可能遇到的情况。例如，研究人员指定他们的研究人群和一些其他的参数，并针对不同的药物和（或）结果进行大量的比较。如果研究者的问题符合通用模板的要求，则有多种自动生成研究人群基本描述和研究方案中所要求的其他指标参数的方法。历史上这些方法由 OMOP 实验产生，该实验试图通过迭代许多不同的研究设计和参数来评估观察性研究设计是否能很好地重现药物与不良事件之间的因果关系。

OHDSI 方法通过利用通用框架和工具可相对简单地实施这些步骤，以支持在研究方案中纳入可行性与研究诊断（参见下方 19.2.4 节）。

19.1.5 研究套件

标准化模板与设计的另一个原因是即使研究人员认为以方案的形式对研究进行了详细的描述，但实际上可能没有指定足够的元素以生成完整的计算机代码来执行研究。OHDSI 框架支持的相关基本原理是以计算机代码的形式生成完全可追溯和可再现的流程，通常被称为“研究套件”。OHDSI 的最佳实践是在 git 环境中记录这样的研究套件。这种研究套件包含代码库的全部参数和版本戳。正如之前提到的，观察性研究经常提出可能影响公共卫生决策或政策制定的问题。因此，在基于研究结果采取任何实质性举措之前，理论上应由不同的研究人员在各种不同环境中进行重复实验。达到此目标的唯一方法是将研究中遇到的每一个细节都明确地列出并完全再现，以避免猜疑或曲解。为了支持这种最佳实践，OHDSI 工具被设计用以协助将文本文档形式的研究方案转化为计算机或机器可读的研究套件。这种架构的折衷之处是，并非所有的用例或自定义分析都可以使用现有的 OHDSI 工具轻松解决。随着不断进步与发展，OHDSI 社区正针对更多用例增加新的功能。社区中的任何人都可以参与其中，并针对全新的用例提出增加相应新功能的建议。

19.1.6 符合通用数据模型（CDM）的数据

进行 OHDSI 研究的前提条件是将观察性数据库转换为 OMOP 通用数据模型（CDM）。所有的 OHDSI 工具与下游分析步骤均假定所用数据符合 CDM 规范（参见第 4 章）。 因此，ETL 流程对特定的数据源进行充分记录（参见第 6 章）也显得很重要，因为该流程可能会产生人为影响或导致不同站点的数据库之间存在差异。引入 OMOP CDM 的目的就是为了消除在不同站点间数据的特异性差别，但这远非完美的解决方案，目前依然有很多挑战等待着社区的同道们来进一步完善提升。因此，无论是在与本单位人员进行合作，还是在与其他单位进行远程协作研究时，相关人员对原始数据转换为 OMOP CDM 的熟悉程度是保证项目顺利进行的极为重要的基础。

除了 CDM 外，OMOP 标准化词表系统（第 5 章）与 OHDSI 架构一样，二者的共同使用是实现各种数据源之间互联互通的关键基础。标准化词表力求在每个词汇域内定义一组标准概念，并将所有其他源词表系统中的相关概念都映射到该词表统一的标准概念上。这样，当要比较两个不同的药物、诊断或诊疗过程的数据库时，将它们都转化为 CDM ，使不同的源词表系统规范统一为相同的词表系统，就可以进行比较了。同时，OMOP 词汇表还包含层次结构，这些层次结构可用于识别特定队列定义的适当代码。我们再次推荐最佳的做法是将您的词表映射转化，并在下游查询中使用 OMOP 标准化词表，以便在将您的数据库 ETL 转变为 OMOP CDM 并使用 OMOP 词表时可以获得最大化的收益。

19.2 详细研究步骤

19.2.1 定义问题

第一步是将您的研究兴趣转化为一个可以通过观察性研究解决的精确问题。假设您是一名临床糖尿病研究人员，您希望探究针对 2 型糖尿病患者（type 2 diabetes mellitus，T2DM）的护理质量问题。您可以将这个大的目标分解为更具体的问题，这些问题可以归类到前述第 7 章中的三类问题之一。

在一项特征刻画研究中，人们可能会问，“在给定的医疗环境中，轻度 T2DM 患者与重度 T2DM 患者的处方做法是否符合当前建议？”这个问题不是比较两种疗法有效性的因果问题，它只是在描述数据库中与现有临床指南相关的处方实践。

也许您也怀疑 T2DM 治疗的处方指南是否最适合特定的患者亚群，例如既诊断为 T2DM 又患有心脏病的患者。这种询问可以转化为 PLE 研究。具体来说，您可以提出一个关于两种不同的 T2DM 药物在预防心血管事件（例如心力衰竭）方面的相对有效性的问题。您可以设计一项研究，以检查两个诊断为 T2DM 合并心脏病的患者群体在服用不同药物时因心力衰竭住院的相对风险。或者，您可能想要开发一个模型来预测哪些患者将从轻度 T2DM 发展到严重 T2DM。这可以被界定为 PLP 问题，提示患者将有高风险发展为严重 T2DM， 给予该患者预警照护。

从纯粹务实的观点出发，定义研究问题还需要评估回答问题所需的方法是否符合 OHDSI 工具集内的可用功能（可使用现有工具解决的问题类型详见第 7 章）。当然，您也可以设计自己的分析工具或修改当前可用的工具来回答其他问题。

19.2.2 查看数据可用性和质量

在提出特定的研究问题之前，建议您查看数据的质量（请参阅第 15 章），并根据填充的字段以及数据涵盖的医疗场景来真正了解特定的观察性医疗数据库。这可以帮助您快速发现阻碍研究可行性的问题。下面，我们将指出一些可能会出现的常见问题。

让我们回到上面的示例中，为轻度 T2DM 进展到重度 T2DM 的患者建立风险预测模型。理想情况

下，可以通过检查糖基化血红蛋白（HbA1c）水平来评估 T2DM 的严重程度，HbA1c 是一项实验室检测，可以反映患者过去 3 个月血糖的平均水平。这些指标可能适用于所有患者，也可能并不适用于所有患者。如果不能适用于所有或者部分患者，则必须考虑是否可以识别并使用其他有关 T2DM 严重程度的临床指标。再者，如果 HbA1c 值仅适用于部分患者，您还需要评估如果只关注这一部分患者是否会导致研究出现不必要的偏倚。有关缺失数据的问题详见第 7 章。

另一个常见的问题是缺乏特定医疗场景的信息。在上述 PLE 案例中，建议的结局是患者因心力衰竭住院。如果给定的数据库没有患者住院信息，则可能需要考虑另一种结局，以评估不同 T2DM 治疗方法的相对有效性。在其他数据库中，可能无法获得患者门诊诊断数据，因此可能需要考虑该队列的设计。

19.2.3 研究人群

定义研究人群是所有研究的基本步骤。在观察性研究中，代表研究群体的一组个体通常被称为队列。入组所需的患者特征由当前临床问题相关的研究人群来确定。一个简单的队列示例：年龄在 18 岁以上且病历中带有 T2DM 诊断代码的患者。该队列的定义包含两个由 “AND” 逻辑关联的标准。通常，队列定义包含更多的标准，这些标准通过更复杂的嵌套布尔逻辑以及时间参数（例如特定的研究时期或患者基线期所需的时间长度）连接在一起。

一组准确的队列定义需要回顾相关的科学文献，并采纳临床和技术专家的意见，他们了解在解释您的数据库以确定适当的入组患者群体时遇到的一些挑战。需要谨记的是，在使用观察性数据时，这些数据不能提供患者的完整病史，而是某个时间段的信息，其保真度会受到数据记录时引入的人为错误和偏倚的影响。给定的患者只能在有限的时间内（即观察期）进行随访。对于给定的数据库或医疗场景以及正在研究的疾病或治疗方法，临床研究人员能够避免最常见的错误。举一个简单的例子，T2DM 患者诊断最常见的问题是 T1DM 患者有时被误诊为 T2DM。由于 T2DM 患者与 T1DM 患者本质上是不同的群体，因此在针对 T2DM 患者的研究中无意间纳入一组 T1DM 患者可能会使结果偏误。为了对 T2DM 队列有一个可靠的定义，可能需要排除仅接受过胰岛素治疗的糖尿病患者，以避免 T1DM 患者错误入组。当然，也有可能研究者只是对病历中具有 T2DM 诊断代码的所有患者的特征感兴趣。在这种情况下，可能不适合采用进一步的入组标准来尝试移除错误编码的 T1DM 患者。

一旦确定了一个或多个研究人群的定义，则可以开始利用 OHDSI 工具 ATLAS 来创建相关队列。第 8 章和第 10 章详细介绍了 ATLAS 和队列人群生成过程。简而言之，ATLAS 提供了一个可以利用详细入组标准来定义和生成队列的用户界面（UI）。在 ATLAS 中定义队列后，用户可以直接以人类可读的格式导出队列的详细定义，并将其加入到研究方案中。如果由于某种原因未能将 ATLAS 实例连接到观察性医疗数据库，仍可以使用 ATLAS 来创建队列定义，并直接导出底层的 SQL 代码并将其合并到研究套件中，以在 SQL 数据库服务器上单独运行。建议尽可能直接使用 ATLAS，因为 ATLAS 除了能为队列定义创建 SQL 代码，还具有其他优点（详见下文）。在极少的情况下，队列定义无法用 ATLAS UI 部署，这时可能需要手动创建 SQL 代码。

ATLAS UI 支持大量可选择的标准来定义队列。可以根据 OMOP CDM 的任何域（例如疾病、药物、操作等）来定义进入和退出队列的标准以及基线标准，必须为每个域指定标准的代码。此外，可以在 ATLAS 中利用基于域的逻辑筛选以及基于时间的筛选来定义研究时段和基准时间窗。为每个标准选择代码时，ATLAS 将会非常有用。ATLAS 集成了词汇浏览功能，可用于构建队列定义所需的代码集。此功能仅依赖于 OMOP 标准词汇表，并且可以选择在词汇表层次结构中包含所有子节点（参见第 5

章)。需要注意的是，此功能要求在 ETL 处理过程中将所有代码都正确映射到标准代码 (参见第 6 章)。如果尚不清楚入组标准应该使用的最佳代码集，则可能需要对队列定义进行一些探索性分析。或者，可以考虑使用更正式的敏感性分析来说明不同代码集对队列的不同定义。

假设 ATLAS 连接数据库配置正确，则在 ATLAS 中可以直接生成定义队列的 SQL 查询语句。ATLAS 将自动为每个队列分配一个唯一 ID，该 ID 可在后端数据库中直接引用队列以供将来使用。可以直接在 ATLAS 中利用队列进行发病率研究，也可以通过 PLE 或 PLP 研究套件中的代码直接指向后端数据库中的队列。对于某个队列，ATLAS 仅保存队列中的患者 ID、索引日期和队列退出日期。这些信息足以推导出患者所有其他属性或协变量，例如用于描述性研究、PLE 和 PLP 研究中的患者基线协变量。

创建队列后，在 ATLAS 中可以直接创建和查看患者人口学数据的基本特征，以及常见药物和疾病的分布情况。

实际上，大多数研究需要指定多个队列或多组队列，然后以各种方式进行比较以获得新的临床结论。对于 PLE 和 PLP 研究，OHDSI 工具提供了一个结构化框架来定义多个队列。例如，在 PLE 相对有效性研究中，通常您至少会定义 3 个队列，一个目标队列、一个比较队列和一个结局队列（参见第 12 章)。此外，要进行完整的 PLE 相对有效性研究，您还需要一些队列作为阴性对照结局和阳性对照结局。OHDSI 工具集提供了快速生成队列的方法，如第 18 章所述，某些情况下这些工具集可以自动生成阴性和阳性对照队列。

最后，OHDSI 社区一些正开展的研究会定义一个可靠的、经过验证的表型库，可导出队列定义为其他研究提供参考。如果现有的队列定义适用于您的研究，则可通过将 json 文件导入 ATLAS 实例来获取确切的队列定义。

19.2.4 可行性和研究诊断

一旦确定和生成队列，就可以通过更正式的步骤来审查现有数据源的研究可行性，并将结果总结在最终的研究方案中。对研究可行性的评估可包括许多探索性以及少量迭代性的工作步骤。下面我们会介绍一些常见的工作步骤。

在此阶段，主要工作是全面检查队列的特征分布，以确保您生成的队列符合所需的临床特征，并标记所有意料之外的特征。回到前述的 T2DM 案例，通过遍历所有诊断的分布来对这个简单的 T2DM 队列进行特征刻画，可能会发现 T1DM 患者入组或其他意料之外的问题。研究方案中最好加入新队列的初始特征刻画这一步骤，该步骤可检查队列定义是否具有临床有效性。在实施方面，想要一次通过运行，最简单的方法就是检查利用 ATLAS 创建队列时自动生成的队列人口学资料、首选用药以及所患疾病等特征。如果在 ATLAS 中直接创建队列的选项不可用，则可以通过手动编写 SQL 或使用 R 特征提取包来刻画队列的特征。实际上，在较大的 PLE 研究或 PLP 研究中，这些步骤可以内置到具有特征提取步骤的研究套件中。

评估 PLE 或 PLP 可行性的另一个常见且重要步骤是评估队列规模以及目标组和比较组的结局数量。ATLAS 的发病率特征可用于查找这些数量，如其他地方所述，这些数量可用于计算研究的执行效力。

对于 PLE 研究，强烈建议完成倾向性评分（propensity score，PS）匹配步骤和相关的研究诊断，以确保目标组和比较组中的人群之间有足够的重叠。这些步骤在第 12 章中作了详细说明。此外，使用这些最终匹配的队列，可以计算统计效力。

在某些情况下，OHDSI 社区只有在开展研究后，才通过在现有可用样本量条件下报告最小可检测

相对风险（minimal detectable relative risk，MDRR）来检验统计效力。当跨多数据库和站点开展高吞吐量、自动化的研究时，此方法可能会更有用，并且在执行所有分析之后，在现有数据库中计算统计效力优于预过滤分析之后计算统计效力。

19.2.5 完成研究方案和研究套件

一旦完成了前面所有步骤之后，就应该封装最终的研究方案，其中应包括详细的队列定义和理论上从 ATLAS 导出的研究设计的信息。在附录 D 中，我们提供了完整的 PLE 研究方案的目录示例表。您也可以在 OHDSI github 上找到。我们提供此示例作为全面的指南和清单，但请注意某些部分可能与您的研究无关。

如图 19.1 所示，在封装研究者可读的最终研究方案的同时，应该准备具有机器可读研究代码的最终研究套件。下图将这些后续的步骤称为研究部署，包括从 ATLAS 导出最终的研究套件和（或）开发可能需要的自定义代码。

已完成的研究套件可用于执行初步的研究诊断步骤，这些步骤可以在研究方案中描述。例如，在一项新的 PLE 研究队列中比较两种治疗的相对有效性，开始执行的研究诊断步骤包括创建队列，以及创建和匹配倾向分数确保目标队列和比较队列有足够的重叠，以满足研究的可行性。一旦确定了这一点，就可以用匹配的目标队列、比较队列和结局队列相交来进行效力计算以获得结局数量，效力计算的结果可以在研究方案中加以描述。根据这些研究诊断结果，可以决定是否继续执行最终结局模型。在描述性或 PLP 研究流程中，在此阶段可能需要完成类似的步骤，但不在此尽述。

尤为重要的是，我们建议在这个阶段让临床合作者和利益相关者对您的最终方案进行审查。

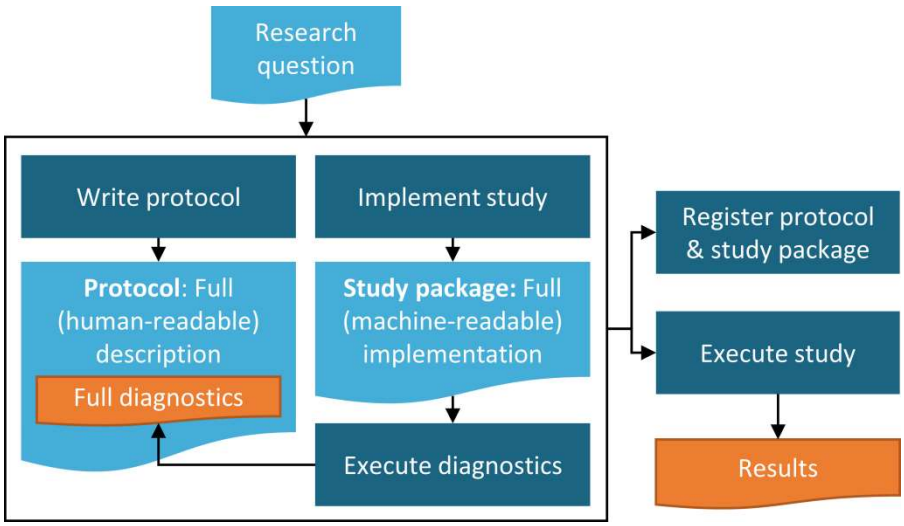


图 19.1：研究流程

19.2.6 执行研究

完成前面所有的步骤后，理论上研究的执行过程很简单。当然，应该检查代码或流程是否符合研究方案中列出的方法和参数。可能还需要测试和调试研究套件，以确保它在您的研究环境中准确运行。

19.2.7 阐述和撰写

一个定义明确的研究，其样本量是足够的，数据质量是合理的，研究结果的解释通常是直截了当的。类似地，除了撰写最终结果外，形成最终报告的大部分工作都在规划和创建研究方案过程中完成了，因此，最终发表的报告或手稿的撰写过程通常也很简单。

但在某些常见情况下，结果的阐述却具挑战性，应谨慎对待。

1. 样本容量可以影响显著性和置信区间；
2. 针对 PLE：用阴性对照进行 p 值校准可能会出现较大的偏差；
3. 在实施研究的过程中，可能会出现意外的数据质量问题。

对于任何给定的研究，由研究者自行决定是否报告上述问题，以及是否对研究结果的解释进行相应的调整。与方案制定过程一样，我们建议在发布最终报告或提交稿件之前，让临床专家和利益相关者对研究结果及其解释进行审查。

19.3 总结

- 研究应该探究一个明确定义的问题。
- 提前进行适当的数据质量、完整度和相关性的检查。
- 建议在研究方案制定时邀请源数据库相关专家。
- 提前在研究方案中记录将要进行的研究。
- 在实施最终的研究前，同时生成研究套件的代码和文字版的研究方案，并对研究进行可行性分析与诊断。
- 在研究实施前需要进行注册，如有审批要求，还需要进行审批。
- 最终的报告或手稿需要通过相关临床专家和其他利益相关者的审核。

第 20 章 OHDSI 协作网研究

章节负责人： Kristin Kostka, Greg Klebanov & Sara Dempster

OHDSI 的任务是通过观察性研究产生高质量的证据。实现这一目标的主要方式是通过协作性研究。在先前的章节中，我们讨论了 OHDSI 社区如何制定标准和工具来协助开展高质量、可重复的研究，包括利用 OMOP 标准化术语表、通用数据模型（CDM）、分析方法套件、ATLAS 和研究步骤（第 19 章）来进行回顾性数据库研究。 OHDSI 协作网研究是一种透明、一致且可重复的高水平研究方法，可对大量地理上分散的数据进行研究。在本章中，我们将讨论 OHDSI 协作网研究由什么构成，如何开展一项协作网研究，以及背后的关键支撑技术，如 ARACHNE Research Network。

20.1 OHDSI 研究协作网

OHDSI 协作网研究致力于推进医疗健康领域观察性研究的国际合作。如今，该网络由 100 多个按照 OMOP 通用数据模型标准化的数据库组成，总共有 10 亿多条患者记录。OHDSI 是一个开放式协作网，邀请全世界拥有患者水平数据的医疗机构加入 OHDSI 协作网，将各自数据转换为 OMOP CDM，并进行协作性研究。当数据转换完成后，将邀请协作者们在 OHDSI 项目经理负责维护的数据网络中同步站点信息。每个 OHDSI 协作网的站点都是自愿加入的，并没有硬性要求。每个站点可选择加入各自的协作网研究。每项研究的数据保留在各个站点的防火墙后。患者水平的数据并不在站点之间流动。各站点间仅共享研究结果。

数据所有者加入 OHDSI 协作网研究的好处

–**使用免费工具**：OHDSI 发布免费的开源工具，用于数据表征和标准化分析（如：浏览临床概念、定义和描述队列，进行群体水平评估和患者水平预测研究）。

–**参与到主要的研究社区中**：撰写和发表协作性研究，与各学科负责人和利益相关者团体紧密合作。

–**提升临床技能水平**：协作网研究可以在数据合作伙伴间实现临床特征刻画和质量提升基准。

20.2 OHDSI 协作网研究

在第 19 章中，我们介绍了使用 CDM 开展研究时通常的注意事项。通常来说，研究者可以在单个 CDM 或多个 CDM 上开展研究。研究可以在单个机构的 CDM 数据中执行，也可以在多个机构中执行。在本节中，我们将讨论为什么您会希望将跨多个机构的分析研究扩展为协作网研究。

20.2.1 开展 OHDSI 协作网研究的原因

观察性研究的典型用例是在“真实世界”中检验治疗的相对有效性或安全性。 具体来说，您可能打算在上市后重复临床试验，以提高临床试验结果的通用性。 或者，由于有的治疗属于非适应证治疗，您可能希望开展一项研究，以比较临床试验中从未比较过的治疗方法。 再或者，您可能想要研究临床试验中罕见的、上市后才出现的不良反应。为了解决这些研究问题，仅使用您所在机构的一个或两个数

据库不足以开展一项观察性研究，因为您的研究结果仅对特定群体患者有意义。

观察性研究的结果可能会受到许多因数据源位置而异的因素的影响，例如依从性、遗传多样性、环境因素或整体健康状况。即便别人对相同问题进行研究，在临床试验中此类因素也可能存在差异。因此，利用协作网进行观察性研究的主要原因就是为了增加数据源的多样性，或者增加潜在的研究人群以了解结果的通用性。换句话说，研究结果可以在多个站点重复，或者它们是否存在差异，如果存在差异，差异可以解释吗？

因此，借助协作网进行研究，可以通过各种各样的研究设计和数据源，来研究“真实世界”因素对观察研究结果的影响。

20.2.2 定义一项 OHDSI 协作网研究

一项研究何时被视为协作网研究？ 在不同机构的多个 CDM 上运行 OHDSI 研究时，该研究可以被视为 OHDSI 协作网研究。

OHDSI 协作网研究方法包括 OMOP CDM 和标准化工具及研究套件，它们完全指定了所有和研究有关的参数。OHDSI 标准化分析是专门为减少人为因素和提高协作网研究效率与可扩展性而设计的。

协作网研究是 OHDSI 研究社区的重要组成部分。但是，社区没有强制要求将一项 OHDSI 研究打包并共享到整个 OHDSI 协作网。您仍然可以在单个机构中使用 OMOP CDM 和 OHDSI 方法库进行研究，或在您指定的几个机构内共享。这些研究贡献对社区同样重要。是否将研究设计在单个数据库上运行，或是在有限的合作伙伴中开展，或是开放到整个 OHDSI 协作网中，完全由每个研究者决定。本章旨在介绍在 OHDSI 社区进行的开放式协作网研究。

开放式 OHDSI 协作网研究的要素：进行开放式 OHDSI 协作网研究时，您将构建完全透明的研究。此类 OHDSI 研究具有独特的要素，包括：

- 所有文档、研究代码和后续结果在 OHDSI GitHub 上公开可用。
- 研究人员必须创建并发布公共研究方案，详细说明要进行研究分析的范围和意图。
- 研究人员必须使用符合 CDM 的代码创建研究套件（通常使用 R 或 SQL）。
- 鼓励研究人员响应 OHDSI Community Calls，以促进和招募合作者开展 OHDSI 协作网研究。
- 研究分析结束时，汇总的研究结果可在 OHDSI GitHub 中获得。
- 在可能的情况下，鼓励研究人员将研究的 R Shiny 应用发布到 data.ohdsi.org。

在下一节中，我们将讨论如何创建自己的协作网研究，以及协作网研究的独特设计原则及保障条件。

20.2.3 OHDSI 协作网研究的设计原则

设计一项可在 OHDSI 协作网上运行的研究，需要在设计和组装代码方面进行范式转换。 通常，您会在设计研究时考虑目标数据集。这样的话，您可以编写与您分析数据集中真实信息相关的代码。例如，如果您要建立血管性水肿队列，可以只选择 CDM 中代表血管性水肿的概念代码。但如果您的数据处于特定的医疗场景（例如初级医疗保健，非住院医疗场景）或特定于某个地区（例如美国），则可能会出现问题。 您的代码选择可能会影响您的队列定义。

在 OHDSI 协作网研究中，您不仅仅为自己的数据设计和构建研究套件，而是一个可以在全球多个站点上运行的研究套件。您永远不会看到其他参与机构的基础数据，OHDSI 协作网研究仅共享结果文

件。您的研究套件只能收集 CDM 域中可用的数据。您需要用详尽的方法来创建概念集，以满足观察性数据涵盖的医疗场景的多样性。 OHDSI 研究套件通常在所有站点上使用相同的队列定义。这意味着您必须从整体上考虑，避免将同类定义偏向网络中合格数据的子集（例如，围绕医保赔付的数据或 EHR 特有的数据）。建议您编写一个详尽的、可以跨多个机构的 CDM 进行移植的队列定义。 OHDSI 研究套件在所有站点上使用相同的参数化代码集，只对连接到数据库层和存储本地结果进行了较小的调整。稍后，我们将讨论从不同数据集解释临床发现的意义。

除了临床编码差异之外，您还需要考虑本地技术基础设施中预期的差异。您的研究代码将不再在单一的技术环境中运行。每个 OHDSI 协作网站点都会对其数据库层做出独立的选择。那么，您无法将研究套件硬编码为特定的数据库。研究代码需要参数化为一种 SQL 类型，可以容易地修改为 SQL 用语的操作符。幸运的是，OHDSI 社区提供了诸如 ATLAS，DatabaseConnector 和 SqlRender 之类的解决方案，使您的研究套件通用化，以实现跨不同数据库用语的 CDM 兼容性。同时鼓励 OHDSI 研究人员向其他协作网研究机构寻求帮助，以测试和验证研究套件在不同环境中的运行能力。 当出现编码错误时，OHDSI 研究人员可以利用 OHDSI 论坛来讨论和调试研究套件。

20.2.4 OHDSI 协作网研究的保障条件

OHDSI 是一个开放的科学研究社区，OHDSI 协调中心提供了社区基础设施，使合作者能够领导和参与社区研究。每个 OHDSI 协作网研究都需要一名首席研究员，可以是 OHDSI 社区中的任何合作者。OHDSI 协作网研究需要在首席研究人员、合作研究人员及网络数据的参与者之间进行协调。每个站点各自都必须尽职调查，以确保研究方案被批准并授权在本地 CDM 上执行。数据分析师可能需要寻求当地 IT 团队的帮助，以获得适当的权限来运行研究。每个站点研究团队的规模和范围取决于研究的规模和复杂性，以及该站点采用 OMOP CDM 和 OHDSI 工具的成熟度。各站点开展 OHDSI 协作网研究的经验水平也会影响其所需人员。

对于每项研究，各站点启动时可能需要完成以下工作：

- 在机构审查委员会（或同等机构）中注册研究
- 获得机构审查委员会的批准以实施研究
- 得到数据库权限以将模式读取/写入到已批准的 CDM
- 确保配置一个功能齐全的 R Studio 环境来执行研究套件
- 检查研究代码中是否存在任何技术异常
- 与本地 IT 团队合作，允许并安装在技术允许范围内的 R 工具包

数据质量和协作网研究：正如在第 6 章中讨论的那样，质量控制是 ETL 过程的基础和重复部分。质量控制应该在协作网研究过程之外定期进行。而对于协作网研究，研究负责人可能会要求查看参与站点的数据质量报告或进行自定义 SQL 查询，以了解数据源中的可能变化。有关 OHDSI 正努力提高数据质量的更多详细工作，请参阅第 15 章。

每个站点都会有一个本地数据分析师来执行研究套件。此人必须检查研究套件的输出，以确保不会传输任何敏感信息。当您使用诸如群体水平效果评估（Population-Level Effect Estimation ，PLE）和患者水平

预测（Patient Level Prediction ，PLP）之类的预先建立的 OHDSI 方法时，对于一个给定的分析，会有最小单元数的可配置设置。数据分析师必须检查这些设置阈，并确保其遵守当地政策。

共享研究结果时，数据分析师必须遵守当地所有政策，包括结果传输方法和外部公开发表的审批流程。OHDSI 协作网研究不共享患者水平数据。换句话说，来自不同站点的患者水平数据不会被集中到一个中心环境。研究套件的创建旨在创建汇总结果的相关文件（例如汇总统计信息、点估计值、诊断图等），并且不共享患者水平的信息。许多组织不需要在参与研究团队成员之间签订数据共享协议。但是，根据所涉及的机构和数据来源，可能有必要制定更正式的数据共享协议，并由特定的研究团队成员签署。如果您是有兴趣参加协作网研究的数据所有者，建议您咨询当地的政府，了解已制定哪些政策，并且必须符合哪些政策才能加入 OHDSI 社区研究。

20.3 开展 OHDSI 协作网研究

开展 OHDSI 协作网研究分为三个步骤：

- 研究设计和可行性分析
- 研究实施
- 研究结果的发布与发表

20.3.1 研究设计和可行性分析

研究可行性分析阶段（或研究前阶段）是指定义一个研究问题，并通过描述研究方案来回答这个问题的过程。这一阶段的重点是评估跨参与站点实施研究方案的可行性。

可行性分析阶段的结果是生成最终的研究方案和研究套件，并发布以供协作网实施。正式方案将详细说明研究团队情况，包括指定研究负责人（通常是出版物的通讯作者），以及确定研究时间节点。研究方案对所添加的网络站点在 CDM 数据中进行完整研究套件的审查、批准和执行十分关键。研究方案必须包括研究人群、使用的方法、如何存储和分析研究结果以及研究结束后如何发布研究结果（例如出版物发表、学术会议展示等）等内容。

可行性分析阶段很难具体定义，它是与研究类型密切相关的一系列工作。至少，研究负责人应该确定各协作网站点目标患者人群相应的药物暴露、临床处理、病情或人口统计等信息。在可能的情况下，研究负责人应该暂时使用自己的 CDM 来设计目标队列。但是，并不要求研究负责人必须使用真实患者数据的实时 OMOP CDM 来开展协作网研究，可以利用合成数据（如 CMS 合成公共使用文件、Mitre 或 Synthea 合成数据等）设计目标队列，并请 OHDSI 各协作网站点的合作者帮助验证该队列的可靠性。可行性分析包括要求合作者使用来自 ATLAS 或测试 R 工具包的定义队列的 JSON 文件，并执行第 19 章中讨论的初步诊断来创建和描述队列。与此同时，研究负责人需要建立组织内特定的流程来批准实施该 OHDSI 研究——例如组织机构审查委员会的批准。研究负责人有责任在可行性分析阶段完成这些组织特定的审查任务。

20.3.2 研究实施

在完成可行性分析阶段后，研究进入实施阶段。在这个阶段，OHDSI 协作 网各站点可以选择参与分析，并重点讨论研究设计和保障条件。

在研究负责人与 OHDSI 社区联系，正式发布开展一项新的 OHDSI 协作网研究，并正式开始招募参与站点时，研究将进入实施阶段。研究负责人将在 OHDSI GitHub 发布研究方案，并在每周的 OHDSI 会议和 OHDSI 论坛上发布该项研究，以邀请中心和合作者参与。当一些站点选择参与时，研究负责人

将直接与每个站点沟通，并提供有关 GitHub 库的信息，该库发布了研究方案和代码，以及如何执行研究套件的说明。理想情况下，协作网研究由所有站点并行实施，因此最终结果可同时共享，以确保所有站点的团队成员可以了解其他团队的研究发现而不会产生偏倚。

针对每个站点，研究小组将确保研究遵循所在机构的程序，以获得参与研究、实施研究和对外分享结果的批准。可能包括获得机构审查委员会（IRB）的批准，或同等机构对指定方案的批准。当研究被批准实施时，各站点的数据科学家/统计学家将按照研究负责人的指示访问 OHDSI 研究套件，并按照 OHDSI 指南标准化格式生成结果。每个参与站点将遵循内部机构有关数据共享规则的流程。除非获得了 IRB 或其他机构的批准，否则各站点不应该共享结果。

研究负责人将负责告知如何接收结果(例如通过 SFTP 或 Amazon S3 bucket)，以及相应的时限。各站点可指出该传输方法是否符合其内部协议，并据此制定替代方法。

在研究实施阶段，如果需要合理的调整，整个研究团队（包括研究负责人和各参与站点团队）可以共同优化研究方案。如果研究方案超出了批准的范围和程度，各参与站点有责任通过与研究负责人协同将情况传达给他们的组织，更新方案后重新提交以供当地 IRB 审查和重新批准。

研究负责人和所有数据科学家/统计学家的最终职责是跨中心汇总结果并进行适当的荟萃分析。OHDSI 社区已经验证了将多个网络站点共享的结果文件聚合为统一结果的方法。EvidenceSynthesis 包是一个可免费获取的 R 工具包，包含跨多个来源（如分布式研究中的多个数据节点）合并证据和诊断的程序，具有荟萃分析和绘制森林图等功能。

研究负责人需要监督各站点参与情况，并对各站点进行定期检查以帮助消除实施方案过程中遇到的障碍。研究的实施不是一成不变的，各协作网站点的运行环境可能存在数据库层（例如访问权/访问模式的许可）或分析工具（例如无法安装所需的包、无法通过 R 访问数据库等）等方面的挑战。各参与研究的站点可自主决策，沟通研究过程中存在的问题，并对是否调用适当资源以协助解决其 CDM 所遇到的问题做最终决定。

虽然 OHDSI 研究可以快速实施，但建议给所有参与站点合理的时间，以实施研究和获得批准发布研究结果。新的 OHDSI 协作网站点可能会出现他们参与的第一个协作网研究比正常情况下要花费更长时间的现象，因为他们要处理数据库权限或分析库更新等环境配置问题。OHDSI 社区在这方面提供了支持，如果遇到问题，可以在 OHDSI 论坛上寻求答案。

研究负责人应在方案中设置研究进程节点，并提前沟通预期的研究结束日期，以协助管理整个研究时间节点。如果没有遵守研究时间节点，研究负责人有责任通知参与站点更新研究计划，并管理研究实施的整体进展。

20.3.3 研究结果的发布与发表

在结果发布和发表阶段，研究负责人将与其他参与者合作完成各种管理任务，如完善研究手稿和优化数据可视化。一旦研究开始实施，研究结果就被集中存储以供研究负责人进一步分析。研究负责人负责创建和发布完整的研究结果（如 Shiny 应用），供各参与者审查。如果研究负责人使用 OHDSI 研究框架，无论是由 Atlas 生成的还是从 GitHub 代码手动修改的，都将自动创建 Shiny 应用。研究负责人如果想创建自定义代码，可以使用 OHDSI 论坛请求帮助，为他们的研究套件创建 Shiny 应用。

不确定在哪里发布您的 OHDSI 协作网研究？请咨询 JANE(Journal/Author Name Estimator)，这是一个可以提取您的摘要来

在撰写手稿时，我们鼓励每个参与的合作者审查并确认手稿符合外部出版流程。至少，参与站点应该指定一个出版物负责人，这个人将确保在手稿准备和提交过程中遵守内部流程。选择哪本期刊来提交研究由研究负责人决定，但应该遵从研究开始时共同讨论的结果。OHDSI 研究的所有合著者都应该符合 ICMJE 作者指南 2。他们可以选择在任何论坛中进行结果的展示（例如 OHDSI 研讨会、其他学术活动或期刊出版物）。研究人员也可以在每周的 OHDSI 会议和全球 OHDSI 研讨会上展示其 OHDSI 协作网研究。

20.4 未来展望：协作网研究自动化

目前协作网研究的过程是需要人工管理的——研究小组成员通过各种方式（包括 Wiki、GitHub 和电子邮件）合作，实现研究设计、代码共享和研究结果共享。这样的研究过程存在不一致性和不可扩展性，为解决这个问题，OHDSI 社区正积极提升，将协作网研究过程系统化。

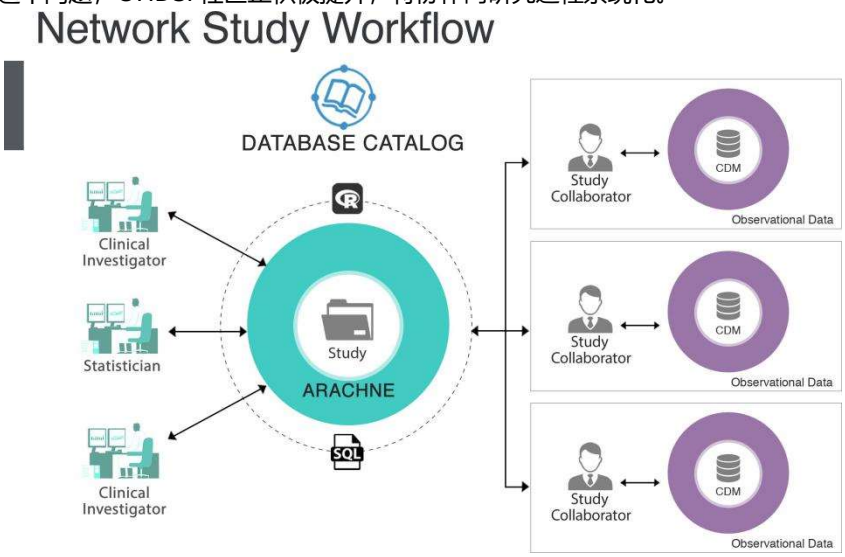


图 20.1： ARACHNE 协作网研究过程

ARACHNE 是一个旨在简化和自动化协作网研究过程的平台。它使用 OHDSI 标准，在多个组织之间建立一致的、透明的、安全的和兼容的观察性研究过程。ARACHNE 对通信协议进行了标准化以实现协作网研究各站点间数据访问和分析结果交换，并对受限制的内容使用身份验证和授权机制。它将参与的组织——数据提供者、调查人员、赞助商和数据科学家——纳入一个单独的协作研究团队，以促进端到端的观察性研究协调。ARACHNE 支持创建一个完整的、基于标准的 R、Python 与 SQL 的执行环境，并包含数据管理员控制的审批工作流。

ARACHNE 构建的目的是实现与其他 OHDSI 工具的无缝集成，包括 ACHILLES 报告，导入 ATLAS 设计工件，以及创建自包含套件并自动跨多个站点执行这些套件。未来的设想是将多个协作网连接在一起，不仅应用于单个协作网中的各组织之间的研究，还可应用于跨多个协作网的各组织之间的研究。

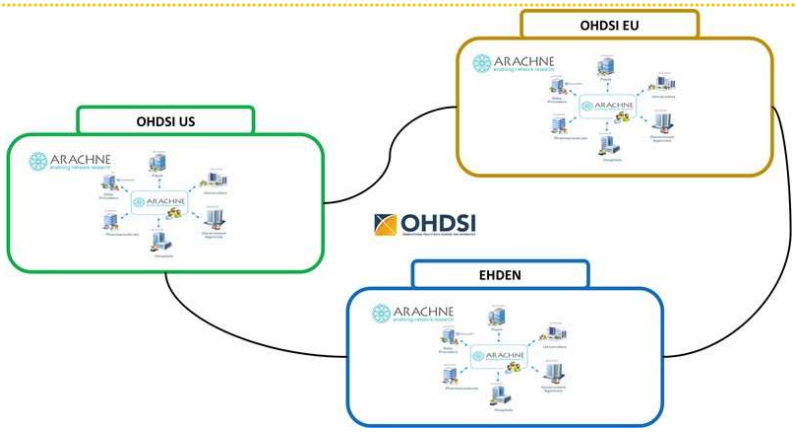


图 20.2： ARACHNE 协作网示意图

20.5 最佳的 OHDSI 协作网研究实践

当您进行协作网研究时，OHDSI 社区可以帮助您确保遵循最佳的 OHDSI 协作网研究实践。

研究设计和可行性：在进行协作网研究时，请确保您的研究设计不偏向单一类型的数据。可能或多或少需要协调队列定义以保证所有站点纳入人群的一致性，这取决于数据类型的异构程度以及研究站点在将数据转换为 OMOP CDM 时严格遵守标准化约定的程度。这一点之所以如此重要，是因为需要控制跨协作网站点的数据获取、表示和转换与临床上数据获取、表示和转换的差异。特别是在比较药物有效性的研究中，确保跨站点间暴露队列和结局队列定义的一致性可能具有挑战。例如，药物暴露信息可以来自不同的数据源，而这些数据源可能因其错误分类的可能性不同而不同。医疗保险计划内的配药才能获得医疗保险赔付，因此药物获得医疗保险赔付时，很有可能填写了医疗保险计划内的处方单。然而，EHR 中纳入的处方订单是任意的，不与其他数据关联，订单可以是医疗保险计划等特定处方，也可以是单纯购买的处方订单。医生开具处方记录的时间、药剂师开具处方的时间、病人到药房取药的时间、以及病人实际服用第一颗药丸的时间之间可能存在时间差，这种测量误差可能会使任何分析用例间的结果产生偏差。因此，在制定研究方案时，进行研究可行性分析以评估纳入数据库的恰当性非常重要。

研究实施：在可能的情况下，建议研究负责人利用 ATLAS、OHDSI Methods Library 和 OHDSI Study Skeletons 工具以尽可能使用标准化分析套件创建研究代码。利用 OHDSI 套件可实现按 CDM 兼容和跨数据库层的方式来创建研究代码。确保所有函数和变量参数化（例如，不对数据库做硬连接，设定本地硬盘驱动器路径，设立一个特定的操作系统）。在招募参与站点时，应确保每个协作网站点符合 CDM 兼容性，并能定期更新 OMOP 标准化术语表。同时对各站点仔细调查，确保每个协作网站点都对其 CDM 进行了数据质量检查并记录在案（例如，确保 ETL 遵守 THEMIS 的业务规则和约定，将正确的数据放入正确的 CDM 表和字段）。建议每个数据分析师在执行研究套件之前，将本地 R 工具包更新为最新版本的 OHDSI 套件。

结果和发布：研究负责人应确保每个站点在共享结果之前遵循本地管理规定。开放的、可重复的科学就是所有的研究设计和实施都是可获取的。OHDSI 协作网研究是完全透明的，所有文档和后续结果都发布到 OHDSI GitHub 库或 data.ohdsi.org R Shiny 服务器上。在准备论文手稿时，研究负责人应该描述 OMOP CDM 原则和标准化术语表，以确保期刊理解数据在 OHDSI 协作网各站点间是如何变化的。例如，如果您正在实施一个使用医疗保险数据库和 EHRs 的协作网研究，期刊审稿人可

能会要求您解释如何跨多个数据类型维护队列定义的完整性。审稿人可能希望了解第 4 章中如何比较 OMOP 观察期与资格文件等内容。其中，资格文件是存在于医疗保险数据库中的文件，该文件对患者是否受保险公司承保进行属性说明。这本质上要求关注数据库本身的人为因素，并关注 ETL（CDM 如何将记录转换为观察）。此时，可以参考 OMOP CDMOBSERVATION PERIOD 如何创建，也可以描述如何使用源系统中的触点创建观察期。在论文手稿讨论部分需要表明 EHR 数据的局限性，不像医疗保险数据覆盖了全时间段内所有支付触点，EHR 数据在患者与医疗服务提供者使用不同的 EHR 记录时没有记录，因此，观察期可能因为患者向 EHR 以外的医疗服务提供者寻求医疗服务而中断。这种使用不同的 EHR 记录是人为设定如何获取系统中数据的差异导致的，并不是临床上存在的差异，但这会使不熟悉 OMOP 如何派生观察期表的人产生困惑。因此，为表明 OMOP 如何派生观察期表，有必要在手稿讨论部分进行解释。类似地，在手稿中描述 OMOP 标准化术语表提供的术语服务也很有用，这可以使各站点获取的临床概念是相同的。将源代码映射到标准概念时总是需要做出决策，此时 THEMIS 公约和 CDM 质量检查可以提供信息，比如信息应该传输到哪里以及数据库在多大程度上遵循了这个原则。

20.6 总结

– 当一项 OHDSI 研究在不同机构的多个 CDM 上运行时就成为一项 OHDSI 协作网研究。

– OHDSI 协作网研究对所有人开放。任何人都可以发起协作网研究。任何与 OMOP 兼容的数据库都可以选择参与研究并贡献研究结果。

– 您需要帮助开展协作网研究吗？请咨询 OHDSI 研究培育委员会，我们可以帮助您设计和开展协作网研究。

– 友情分享。所有的研究文档、代码和结果都发布在 OHDSI GitHub 或 R Shiny 应用程序中。同时，在 OHDSI 活动上也展示了各协作网研究负责人发布的研究成果。

64. http://jane.biosemantics.org/

65. http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

附录

A 术语表

ACHILLES

A database-level characterization report.
数据库水平特征刻画报告。

ARACHNE

The OHDSI platform that is being developed to allow the orchestration and execution of federated network studies.
正在开发的 OHDSI 平台，能够编排和执行协作联邦网络研究。

ATLAS

A web-based application that is installed on participating sites to support the design and execution of observational analyses to generate real world evidence from patient level clinical data.
一个网络应用程序，通过安装在参与单位/机构来支持观察性分析的设计和执行，从而从患者层面的临床数据生成真实世界的证据。

Bias（偏倚）

The expected value of the error (the difference between the true value and the estimated value).
误差的期望值(真实值与估计值之间的差值)

Boolean（布尔运算）

Variable that has only two values (true or false).
仅有两个值(真或假)的变量。

Care site（医疗单位）

A uniquely identified institutional (physical or organizational) unit where healthcare delivery is practiced (office, ward, hospital, clinic, etc.).
一个唯一确定的机构(物理或组织)单元，用于提供医疗保健服务（办公室、病房、医院、诊所等)。

Case control（病例对照）

A type of retrospective study design for population-level effect estimation. Case-control studies match “cases” with the target outcome to “controls” without the target outcome. Then they look back in time and compare the odds of exposure in the cases and the controls.

一种用于群体水平效果评估的回顾性研究设计。病例对照研究将有目标结果的病例与没有目标结果的“对照组”相匹配。然后回顾过去，比较病例组和对照组的暴露比值。

Causal effect（因果效应）

What population-level estimation concerns itself with. One definition equates a “causal effect” as the average of the “unit-level causal effects” in a target population. The unit-level causal effect is the contrast between the outcome had an individual been exposed and the outcome had that individual not been exposed (or been exposed to A as against B).

群体水平效果评估所关心的问题。有的定义将“因果效应”等同于目标人群中“单位水平因果效应”的平均值。单位水平的因果关系是个体暴露与未暴露的结果之间的对比(或暴露于 A 而非 B)。

Characterization（特性描述）

Descriptive study of a cohort or entire database. See Chapter 11.

对一个队列或完整数据库的描述性研究。请参阅第 11 章。

Claims data（索赔数据）

Data generated for the purpose of billing a health insurance company.

为向医疗保险公司索要费用而开发票的数据。

Clinical trial（临床试验）

Interventional clinical study.

干预性临床研究。

Cohort（队列）

A set of persons who satisfy one or more inclusion criteria for a duration of time. See Chapter 10.

在一段时间内满足一个或多个入选标准的一组人员。请参阅第 10 章。

Concept（概念）

A term (with a code) defined in a medical terminology (e.g., SNOMED CT). See Chapter 5.

在医学术语(如 SNOMED CT)中定义的术语(有代码表示)。请参阅第 5 章。

Concept set（概念集）

A concept set is an expression representing a list of concepts that can be used as a reusable component in various analyses. See Chapter 10.

概念集是表示概念列表的表达式，这些概念可在各种分析中用作可重用组件。请参阅第 10 章。

Common Data Model (CDM) （通用数据模型（CDM））

A convention for representing healthcare data that allows portability of analysis (the same analysis unmodified can be executed on multiple datasets). See Chapter 4.

表示医疗健康数据的常规约定，允许分析的移植性(可以在多个数据集上执行同一个未修改的分析)。请参阅第 4 章。

Comparative Effectiveness（相对有效性）

A comparison of the effects of two different exposures on an outcome of interest. See Chapter 12.

比较两种不同的暴露对结果的影响。请参阅第 12 章。

Condition（状况）

A diagnosis, a sign, or a symptom, which is either observed by a provider or reported by the patient.

诊断、体征或症状，由提供者观察到或由患者报告。

Confounding（混杂）

Confounding is a distortion (inaccuracy) in the estimated measure of association that occurs when the primary exposure of interest is mixed up with some other factor that is associated with the outcome.

混杂是一种当研究主要关心的暴露与结果相关的其他一些因素混合在一起时发生的估计关联度量的曲解(不准确)。

Covariate（协变量）

Data element (e.g., weight) that is used in a statistical model as independent variable.

在统计模型中作为自变量使用的数据元素(如：重量)。

Data quality（数据质量）

The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.

使数据适合于特定用途的完整性、有效性、一致性、及时性和准确性的状态。

Device（器械）

A foreign physical object or instrument which is used for diagnostic or therapeutic purposes through a mechanism beyond chemical action. Devices include implantable objects (e.g. pacemakers, stents, artificial joints), medical equipment and supplies (e.g. bandages, crutches, syringes), other instruments used in medical procedures (e.g. sutures, defibrillators) and material used in clinical care (e.g. adhesives, body material, dental material, surgical material).

通过化学作用以外的机制用于诊断或治疗目的的外来实物或仪器。设备包括植入物(例如：心脏起搏器、支架、人工关节)、医疗设备和用品(例如：绷带、拐杖、注射器)、医疗程序中使用的其他仪器(例

如缝合线、除颤器)和临床护理中使用的材料(例如：粘合剂、身体材料、牙科材料、外科材料)。

Drug（药物）

A Drug is a biochemical substance formulated in such a way that when administered to a Person it will exert a certain physiological effect. Drugs include prescription and over-the counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs.

药物是一种按配方制造的生化物质，当给人服用时，会产生一定的生理作用。药物包括处方药、非处方药、疫苗和大分子生物治疗药物。在当地摄入或使用的放射材料不算作药物。

Domain（域）

A Domain defines the set of allowable Concepts for the standardized fields in the CDM tables. For example, the "Condition" Domain contains Concepts that describe a condition of a patient, and these Concepts can only be stored in the condition_concept_id field of the CONDITION_OCCURRENCE and CONDITION_ERA tables.

域为 CDM 表中的标准化字段定义一组允许的概念。

例如，"条件" 域包含描述患者条件的概念，这些概念只能存储在 CONDITION_OCCURRENCE 和 CONDITION_ERA 表的 condition_concept_id 字段中。

Electronic Health Record (EHR) （电子健康档案（HER)）

Data generated during course of care and recorded in an electronic system.

在护理过程中产生并记录在电子系统中的数据。

Epidemiology（流行病学）

The study of the distribution, patterns and determinants of health and disease conditions in defined populations.

对特定人群中健康和疾病状况的分布、模式和决定因素的研究。

Evidence-based medicine（循证医学）

The use of empirical and scientific evidence in making decisions about the care of individual patients.

在决定对个别病人的护理时使用经验和科学证据。

ETL (Extract-Transform-Load) ETL(提取-转换-加载)

The process of converting data from one format to another, for example from a source format to the CDM. See Chapter 6.

将数据从一种格式转换为另一种格式的过程，例如从源格式转换为 CDM。请参阅第 6 章。

Matching（匹配）

Many population-level effect estimation approaches attempt to identify the causal effects

of exposures by comparing outcomes in exposed patients to those same outcomes in unexposed patients (or exposed to A versus B). Since these two patient groups might differ in ways other than exposure, "matching" attempts to create exposed and unexposed patient groups that are as similar as possible at least with respect to measured patient characteristics.

许多群体效应估计的方法试图通过比较具有相同结果的暴露患者与未暴露患者（或比较暴露于 A 与 B)来识别风险因素暴露的因果效应。因为这两个患者组除了不同的暴露方式以外,可能仍具有其他不同之处，"匹配"尝试创建特点尽可能相似的暴露和非暴露的患者组组。

Measurement（测量）

A structured value (numerical or categorical) obtained through systematic and standardized examination or testing of a person or person's sample.

通过对一个人或一个人的样本进行系统和标准化的检验或测试而获得的结构化值(数值的或分类的)。

Measurement error（测量误差）

Occurs when a recorded measurement (e.g., blood pressure, patient age, duration of treatment) differs from the corresponding true measurement.

当记录的测量值(如：患者血压、年龄、治疗时间)与相应的真实测量值不同时，产生测量误差。

Metadata（元数据）

A set of data that describes and gives information about other data and includes descriptive metadata, structural metadata, administrative metadata, reference metadata and statistical metadata.

描述和提供关于其他数据的信息的一组数据，包括描述性元数据、结构元数据、管理元数据、引用元数据和统计元数据。

Methods Library（方法库）

A set of R packages developed by the OHDSI community for performing observational studies.

由 OHDSI 社区开发的一套用于进行观察性研究的 R 包。

Model misspecification（模型误定）

Many OHDSI methods employ statistical models such as proportional hazards regression or random forests. Insofar as the mechanism that generated the data deviate from the assumed model, the model is "mis specified."

许多 OHDSI 方法采用统计模型，如比例风险回归或随机森林模型，如果生成数据的机制偏离了假定的模型，则模型是"错定的/误定的"。

Negative control（阴性对照）

An exposure-outcome pair where the exposure is believed to not cause or prevent the outcome. Can be used to assess whether effect estimation methods produce results in line with the truth. See Chapter 18.

一组暴露-结果实验对，该暴露不会导致或防止结果的发生。可用于评估效果估计方法是否产生与事实相符的结果。请参阅第 18 章。

Observation（观察）

A clinical fact about a Person obtained in the context of examination, questioning or a procedure.

通过检查、询问或临床过程等背景信息获得的关于一个人的临床事实。

Observation period（观察期）

The span of time for which a person is at-risk to have clinical events recorded within the source systems, even if no events in fact are recorded (healthy patient with no healthcare interactions).

在源系统中记录的一个人发生临床事件的风险的时间跨度，即使实际上没有记录任何事件的发生(没有医疗保健交互的健康患者)。

Observational study（观察性研究）

A study where the researcher has no control over the intervention.

研究者无法通过干预进行控制的研究。

OHDSI SQL

A SQL dialect that can be automatically translated to various other SQL dialects using the SqlRender R package. OHDSI SQL is mostly a subset of SQL Server SQL, but allows for additional parameterization. See Chapter 9.

OHDSI SQL 是可以使用 SqlRender R 包自动转换成其他各种 SQL 用语的 SQL 用语。OHDSI SQL 主要是 SQL Server SQL 的一个子集，但是允许附加的参数化。

请参阅第 9 章。

Open science（开放科学）

The movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional. See Chapter 3.

使无论其是业余的还是专业的科学研究(包括出版物、数据、实物样品和软件)及其传播能够被各个层次的具有求知欲的社会民众所接受的活动，请参阅第 3 章。

Outcome（结果）

An observation that provides a focal point for an analysis. For example, a patient-level predictive model might predict the outcome “stroke.” Or a population-level estimation might estimate the causal effect of a drug on the outcome “headache.”

一种分析研究提供焦点的观察。例如，患者水平的预测模型可能预测结果“中风”。 或者，群体水平的评估可能会评估一种药物对结果“头痛”的因果影响。

Patient-level prediction（患者水平预测）

Development and application of predictive models to produce patient-specific probabilities for experiencing some future outcome based on baseline characteristics.

预测模型的开发和应用，根据基线特征而生成特定患者未来经历某些结果的概率。

Phenotype（表型）

A description of physical characteristics. This includes visible characteristics like your weight and hair color, but also your overall health, your disease history, and your behavior.

对物理特征的描述。这包括你的体重、头发颜色等可见特征，也包括你的整体健康状况、疾病史和行为。

Population-level estimation（群体水平评估）

A study into causal effects. Estimates an average (population-level) effect size.

因果关系的研究。估计平均(群体水平)效应大小。

Positive control（阳性对照）

An exposure-outcome pair where the exposure is believed to cause or prevent the outcome. Can be used to assess whether effect estimation methods produce results in line with the truth. See Chapter 18.

一组暴露-结果实验对，该暴露能够导致或防止结果的发生。可用于评估效果估计方法是否产生与事实相符的结果。请参阅第 18 章。

Procedure（过程）

Activity or process ordered by, or carried out by, a healthcare provider on the patient to have a diagnostic or therapeutic purpose.

由医疗服务提供者为诊断或治疗目的而进行的活动或过程。

Propensity score (PS) （倾向性评分(PS)）

a single metric used in population-level estimation to balance populations in order to mimic randomization between two treatment groups in an observational study. The PS

represents the probability of a patient receiving a treatment of interest as a function of a set of observed baseline covariates. It is most often calculated using a logistic regression model where the binary outcome is set to one for the group receiving the target treatment of interest and to zero for the comparator treatment. See Chapter 12.

在一个观察性研究中，为了模拟两个治疗组之间的随机化，人口水平估计中用来平衡群体的单一度量。PS 表示患者接受感兴趣治疗的概率，它是一组观察到的基线协变量的函数。它通常使用逻辑回归模型计算，其中二元结果对于接受目标治疗的相关组设置为 1，对于比较治疗设置为 0。请参阅第 12 章。

Protocol（诊疗方案）

A human readable document that fully specifies the design of a study.

一种人类可读的文档，详细说明了研究的设计。

Rabbit-in-a-Hat

An interactive software tool to help define the ETL from source format to CDM. Uses the database profile generated by White Rabbit as input. See Chapter 7.

一个交互式软件工具，帮助定义从源格式到 CDM 的 ETL。

使用由 White Rabbit 生成的数据库配置文件作为输入。请参阅第 7 章。

Selection bias（选择偏倚）

A bias that occurs when the set of patients in your data deviates from the patients in the population in ways that distort statistical analyses.

当你的数据中的一组患者与总体中的患者不一致时，就会产生偏差，从而使统计分析出现误差。

Self-controlled designs（自控设计）

Study designs that compare outcomes during different exposures within the same patient.

比较同一病人不同的暴露结果的研究设计。

Sensitivity analysis（敏感性分析）

A variant of the main analysis used in a study to assess the impact of an

analysis choice over which uncertainty exists.

一种主要分析方法的变体，用于评估分析选择对不确定性存在的影响。

SNOMED

A systematically organized computer processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting.

一个系统组织的计算机可处理的医学术语集，提供临床文献和报告中使用的代码、术语、同义词和定义。

Study diagnostics（研究诊断）

Set of analytical steps where the goal is to determine whether a given analytical approach can be used (is valid) for answering a given research question. See Chapter 18.

一组分析步骤，目标是确定给定的分析方法是否可以用来合理回答一个给定的研究问题。请参阅第 18 章。

Study package（研究套件）

A computer-executable program that fully executes the study. See Chapter 17.

完全执行研究的计算机可执行程序。请参阅第 17 章。

Source code（源代码）

A code used in a source database. For example an ICD-10 code.

源数据库中使用的代码。例如 ICD-10 代码。

Standard Concept（标准概念）

A concept that is designated as valid concept and allowed to appear in the CDM.

一个被指定为有效概念并允许出现在 CDM 中的概念。

THEMIS

OHDSI workgroup that addresses target data format that is of higher granularity and detail with respect to CDM model specifications.

处理与 CDM 模型规范相关的更高粒度和细节的目标数据格式的 OHDSI 工作组。

Visit（就诊）

The span of time a person continuously receives medical services from one or more providers at a care site in a given setting within the health care system.

一个人从一个或多个医疗服务提供者那里连续获得医疗服务的时长。

Vocabulary（术语集）

A list of words and often phrases, usually arranged alphabetically and defined or translated. See Chapter 5.

单词和常用短语列表，通常按字母顺序排列并定义或翻译。请参阅第 5 章。

White Rabbit

A software tool for profiling a database before defining the ETL to the CDM. See Chapter 6.

在将 ETL 定义为 CDM 之前分析数据库的软件工具。请参阅第 6 章。

使用 ACE 抑制剂的结束日期(表 B.1)

·暴露间隔为 30 天

·暴露结束后增加 0 天

队列撤销策略

按时间撤销队列，间隔大小为 30 天。

概念集定义

B 队列定义

这个附录包含了贯穿全书的队列定义。

2.B 血管紧张素转换酶抑制剂

初始事件队列

有下列任何一种情况的人:

病史中首次出现 ACE 抑制剂的药物暴露（表 B.1）。

并在事件索引日期之前至少 365 天和之后 0 天连续观察，将初始事件限制为：每人所有事件。

将合格队列限制为：每人所有事件。

结束日期策略

自定义药物期退出标准。该策略根据在特定概念集中发现的代码创建一个药物期。如果在一个期限内发现了索引事件，队列结束日期将使用该日期的结束日期。否则，它将使用包含索引事件的观察期结束日期。

表 B.1: 血管紧张素转换酶抑制剂

概念 ID	概念名称	排除	后代	映射
1308216	利诺普利	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1331235	奎那普利	NO	YES	NO
1334456	雷米普利	NO	YES	NO

1335471	贝那普利	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	曲多普利	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1373225	培哚普利	NO	YES	NO

B.2 ACE 抑制剂单药治疗的新使用者

初始事件队列

有下列任何一种情况的人:

首次接触 ACE 抑制剂的药物(表 B.2)，

在事件索引日期之前至少 365 天和之后 0 天连续观察，并限制初始事件:每个人最早的事件。

纳入规则

纳入标准 1：在治疗前 1 年被诊断为高血压。

符合以下所有标准:

至少发生 1 例高血压病(表 B.3)，发病时间为索引开始日期前 365 天至后 0 天。

纳入标准 2：既往无抗高血压药物接触史。

符合下列标准:

高血压药物的药物暴露正好为 0 次(表 B.4)，事件开始于索引开始日期之前的所有天到索引开始日期之前的 1 天之间。

纳入标准 3：是否只采用 ACE 作为单一治疗，且没有相应的联合治疗

具备下列各项条件:

高血压药物的药物时代(表 B.4)在索引起始日期前 0 天到后 7 天之间发生 1 次。

限制符合资格的队列:每个人最早的事件。

结束日期策略

自定义药物期退出标准。该策略根据在特定概念集中发现的代码创建一个药物时期。如果在一个时期内发现了索引事件，队列结束日期将使用该时期的结束日期。否则，它将使用包含索引事件的观察期

结束日期。

使用 ACE 抑制剂的结束日期(表 B.2)

·暴露间隔为 30 天

·暴露结束后增加 0 天

队列撤销策略

按时间撤销队列，间隔大小为 0 天。

概念集定义

表 B.2: 血管紧张素转换酶抑制剂

概念 ID	概念名称	排除	后代	映射
1308216	利诺普利	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1331235	奎那普利	NO	YES	NO
1334456	雷米普利	NO	YES	NO
1335471	贝那普利	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	曲多普利	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1373225	培哚普利	NO	YES	NO

表 B.3: 高血压疾病

概念 ID	概念名称	排除	后代	映射
316866	高血压疾病	NO	YES	NO

表 B.4: 高血压药物

概念 ID	概念名称	排除	后代	映射
904542	氨苯蝶啶	NO	YES	NO
907013	美托拉宗	NO	YES	NO
932745	布美他尼	NO	YES	NO
942350	托拉塞米	NO	YES	NO
956874	呋塞米	NO	YES	NO
970250	螺内酯	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
991382	阿米洛利	NO	YES	NO
1305447	甲基多巴	NO	YES	NO
1307046	美托洛尔	NO	YES	NO
1307863	维拉帕米	NO	YES	NO
1308216	利诺普利	NO	YES	NO
1308842	缬沙坦	NO	YES	NO
1309068	米诺地尔	NO	YES	NO
1309799	依普利农	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1313200	纳多洛尔	NO	YES	NO

1314002	阿替洛尔	NO	YES	NO
1314577	奈必洛尔	NO	YES	NO
1317640	替米沙坦	NO	YES	NO
1317967	阿利吉伦	NO	YES	NO
1318137	尼卡地平	NO	YES	NO
1318853	硝苯地平	NO	YES	NO
1319880	尼索地平	NO	YES	NO
1319998	醋丁洛尔	NO	YES	NO
1322081	倍他洛尔	NO	YES	NO
1326012	依拉地平	NO	YES	NO
1327978	喷布洛尔	NO	YES	NO
1328165	地尔硫卓	NO	YES	NO
1331235	喹那普利	NO	YES	NO
1332418	氨氯地平	NO	YES	NO
1334456	雷米普利	NO	YES	NO
1335471	苯那普利	NO	YES	NO
1338005	比索洛尔	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341238	特拉唑嗪	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	群多普利	NO	YES	NO

1344965	胍法辛	NO	YES	NO
1345858	吲哚洛尔	NO	YES	NO
1346686	依普沙坦	NO	YES	NO
1346823	卡维地洛	NO	YES	NO
1347384	厄贝沙坦	NO	YES	NO
1350489	哌唑嗪	NO	YES	NO
1351557	坎地沙坦	NO	YES	NO
1353766	普萘洛尔	NO	YES	NO
1353776	非洛地平	NO	YES	NO
1363053	多沙唑嗪	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1367500	氯沙坦	NO	YES	NO
1373225	培哚普利	NO	YES	NO
1373928	肼酞嗪	NO	YES	NO
1386957	拉贝洛尔	NO	YES	NO
1395058	氯噻酮	NO	YES	NO
1398937	可乐定	NO	YES	NO
40226742	奥美沙坦	NO	YES	NO
40235485	阿齐沙坦	NO	YES	NO

B.3 急性心肌梗死

初始事件队列

有下列任何一种情况的人:

急性心肌梗死的发生情况(表 B.5)

在事件索引日期之前至少 0 天和之后 0 天连续观察，并限制初始事件:每个人的所有事件。

对于符合主要事件的人，包括:具有以下任何一个标准:

至少发生 1 次住院或急诊入院 (表 B.6)，其中事件开始于索引开始日期之前的所有天和索引开始日期之后的 0 天之间，事件结束于索引开始日期之前的 0 天和索引开始日期之后的所有天之间。

限制初始事件队列:每个人的所有事件。

限制合格队列:所有事件每个人。

结束日期策略

日期偏差退出标准。队列定义的结束日期将是索引事件的开始日期加上 7 天

队列撤销策略

撤销队列按时间划分，间隔大小为 180 天。

概念集定义

表 B.5: 住院或急诊

概念 ID	概念名称	排除	后代	映射
314666	陈旧性心肌梗死	YES	YES	NO
4329847	心肌梗死	NO	YES	NO

表 B.6: 住院或急诊

概念 ID	概念名称	排除	后代	映射
262	急诊室和住院病人访问	NO	YES	NO
9201	住院病人访问	NO	YES	NO
9203	急诊室访问	NO	YES	NO

B.4 血管性水肿

初始事件队列

有下列任何一种情况的人:

- 血管性水肿的情况(表 B.7)，

在事件索引日期之前至少 0 天和之后 0 天连续观察，并限制初始事件:每个人的所有事件。

对于符合主要事件的人，包括:具有以下任何一个标准:

- 至少发生 1 次住院或急诊就诊(表 2.8)，

其中事件开始于索引开始日期之前的所有天和索引开始日期之后的 0 天之间，事件结束于索引开始日期之前的 0 天和索引开始日期之后的所有天之间。

限制初始事件队列:每个人的所有事件。

结束日期策略

这个队列定义的结束日期将是索引事件的开始日期加上 7 天

队列撤销策略

按时间撤销队列，间隔大小为 30 天。

概念集定义

表 B.7: 血管性水肿

概念 ID	概念名称	排除	后代	映射
432791	血管性水肿	NO	YES	NO

表 B.8: 住院或急诊

概念 ID	概念名称	排除	后代	映射
262	急诊室和住院病人访问	NO	YES	NO
9201	住院病人访问	NO	YES	NO
9203	急诊室访问	NO	YES	NO

B.5 新使用噻嗪类利尿剂单药治疗

初始事件队列

有下列任何一种情况的人:

- 首次接触噻嗪类或噻嗪类利尿剂(表 B.9)

在事件索引日期之前至少 365 天和之后 0 天连续观察，并限制初始事件:最早的事件每个人。

纳入规则:

纳入标准 1：在治疗前 1 年被诊断为高血压，并符合以下所有标准:

- 至少发生 1 例高血压疾病(表 B.10)，发生时间为检索日期开始前 365 天至后 0 天。

纳入标准 2：既往无抗高血压药物接触史，符合下列标准:

- 高血压药物的药物暴露正好为 0 次(表 B.11)，事件开始于索引开始日期之前的所有天到索引开始日期之前的 1 天之间。

纳入标准 3：是否只采用 ACE 作为单一治疗，而没有相应的联合治疗。具备下列各项条件:

- 高血压药物的药物时期(表 B.11)在索引起始日期前 0 天到后 7 天之间发生的 1 个完全不同的事件(表 B.11)

限制符合资格的队列:每个人最开始的事件。

结束日期策略

自定义药物时期退出标准。该策略根据在特定概念集中发现的代码创建一个药物时期。如果在一个时期内发现了索引事件，队列结束日期将使用该时期的结束日期。否则，它将使用包含索引事件的观察期结束日期。

使用噻嗪类或噻嗪类利尿剂的使用期限(表 B.9)

- 曝光间隔为 30 天
- 曝光结束后增加 0 天

队列撤销策略

按时间撤销队列，间隔大小为 0 天。

概念集定义

表 B.9: 噻嗪类或噻嗪类利尿剂

概念 ID	概念名称	排除	后代	映射
907013	美托拉宗	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
1395058	氯噻酮	NO	YES	NO

表 B.10: 高血压疾病

概念 ID	概念名称	排除	后代	映射
316866	高血压疾病	NO	YES	NO

表 B.11: 高血压药物

概念 ID	概念名称	排除	后代	映射
904542	氨苯蝶啶	NO	YES	NO
907013	美托拉宗	NO	YES	NO
932745	布美他尼	NO	YES	NO
942350	托拉塞米	NO	YES	NO
956874	呋塞米	NO	YES	NO
970250	螺内酯	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
991382	阿米洛利	NO	YES	NO
1305447	甲基多巴	NO	YES	NO
1307046	美托洛尔	NO	YES	NO
1307863	维拉帕米	NO	YES	NO
1308216	利诺普利	NO	YES	NO
1308842	缬沙坦	NO	YES	NO
1309068	米诺地尔	NO	YES	NO
1309799	依普利农	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1313200	纳多洛尔	NO	YES	NO
1314002	阿替洛尔	NO	YES	NO

1314577	奈必洛尔	NO	YES	NO
1317640	替米沙坦	NO	YES	NO
1317967	阿利吉伦	NO	YES	NO
1318137	尼卡地平	NO	YES	NO
1318853	硝苯地平	NO	YES	NO
1319880	尼索地平	NO	YES	NO
1319998	醋丁洛尔	NO	YES	NO
1322081	倍他洛尔	NO	YES	NO
1326012	依拉地平	NO	YES	NO
1327978	喷布洛尔	NO	YES	NO
1328165	地尔硫卓	NO	YES	NO
1331235	喹那普利	NO	YES	NO
1332418	氨氯地平	NO	YES	NO
1334456	雷米普利	NO	YES	NO
1335471	苯那普利	NO	YES	NO
1338005	比索洛尔	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341238	特拉唑嗪	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	群多普利	NO	YES	NO
1344965	胍法辛	NO	YES	NO
1345858	吲哚洛尔	NO	YES	NO
1346686	依普沙坦	NO	YES	NO
1346823	卡维地洛	NO	YES	NO
1347384	厄贝沙坦	NO	YES	NO
1350489	哌唑嗪	NO	YES	NO
1351557	坎地沙坦	NO	YES	NO
1353766	普萘洛尔	NO	YES	NO
1353776	非洛地平	NO	YES	NO
1363053	多沙唑嗪	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1367500	氯沙坦	NO	YES	NO
1373225	培哚普利	NO	YES	NO
1373928	胼酞嗪	NO	YES	NO
1386957	拉贝洛尔	NO	YES	NO
1395058	氯噻酮	NO	YES	NO
1398937	可乐定	NO	YES	NO

40226742	奥美沙坦	NO	YES	NO
40235485	阿齐沙坦	NO	YES	NO

B.6 开始接受第一线治疗的患者

初始事件队列

有下列任何一种情况的人:

• 首次接触一线高血压药物(表 B.12)

在事件索引日期之前至少 365 天和之后 365 天连续观察，并限制初始事件为:每个人最早的事件。

纳入规则:

具备下列各项条件:

• 在索引开始日期之前的所有天和索引开始日期之前的 1 天之间，高血压药物的药物暴露事件正好为 0 次(表 B.13)

• 至少有 1 例高血压病(表 B.14)，发病时间在指标开始日期前 365 天至后 0 天

限制初始事件队列:每个人最早的事件。

限制符合资格的队列:每个人最早的事件。

结束日期策略:

没有确定的结束日期。默认情况下，队列结束日期将是包含索引事件的观测期的结束。

队列撤销策略

按时间撤销队列，间隔大小为 0 天。

概念集定义

表 B.12: 一线高血压药物

概念 ID	概念名称	排除	后代	映射
907013	美托拉宗	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
1307863	维拉帕米	NO	YES	NO
1308216	赖诺普利	NO	YES	NO
1308842	缬沙坦	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1317640	替米沙坦	NO	YES	NO
1318137	尼卡地平	NO	YES	NO
1318853	硝苯地平	NO	YES	NO
1319880	尼索地平	NO	YES	NO
1326012	依拉地平	NO	YES	NO
1328165	地尔硫卓	NO	YES	NO
1331235	喹那普利	NO	YES	NO
1332418	氨氯地平	NO	YES	NO

1334456	雷米普利	NO	YES	NO
1335471	贝那普利	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	群多普利	NO	YES	NO
1346686	依普沙坦	NO	YES	NO
1347384	厄贝沙坦	NO	YES	NO
1351557	坎地沙坦	NO	YES	NO
1353776	非洛地平	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1367500	氯沙坦	NO	YES	NO
1373225	培哚普利	NO	YES	NO
1395058	氯噻酮	NO	YES	NO
40226742	奥美沙坦	NO	YES	NO
40235485	阿齐沙坦	NO	YES	NO

表 B.13: 高血压药物

概念 ID	概念名称	排除	后代	映射
904542	氨苯蝶啶	NO	YES	NO
907013	美托拉宗	NO	YES	NO
932745	布美他尼	NO	YES	NO
942350	托拉塞米	NO	YES	NO
956874	呋塞米	NO	YES	NO
970250	螺内酯	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
991382	阿米洛利	NO	YES	NO
1305447	甲基多巴	NO	YES	NO
1307046	美托洛尔	NO	YES	NO
1307863	维拉帕米	NO	YES	NO
1308216	利诺普利	NO	YES	NO
1308842	缬沙坦	NO	YES	NO
1309068	米诺地尔	NO	YES	NO
1309799	依普利农	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1313200	纳多洛尔	NO	YES	NO
1314002	阿替洛尔	NO	YES	NO

1314577	奈必洛尔	NO	YES	NO
1317640	替米沙坦	NO	YES	NO
1317967	阿利吉伦	NO	YES	NO
1318137	尼卡地平	NO	YES	NO
1318853	硝苯地平	NO	YES	NO
1319880	尼索地平	NO	YES	NO
1319998	醋丁洛尔	NO	YES	NO
1322081	倍他洛尔	NO	YES	NO
1326012	依拉地平	NO	YES	NO
1327978	喷布洛尔	NO	YES	NO
1328165	地尔硫卓	NO	YES	NO
1331235	喹那普利	NO	YES	NO
1332418	氨氯地平	NO	YES	NO
1334456	雷米普利	NO	YES	NO
1335471	苯那普利	NO	YES	NO
1338005	比索洛尔	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341238	特拉唑嗪	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	群多普利	NO	YES	NO
1344965	胍法辛	NO	YES	NO
1345858	吲哚洛尔	NO	YES	NO
1346686	依普沙坦	NO	YES	NO
1346823	卡维地洛	NO	YES	NO
1347384	厄贝沙坦	NO	YES	NO
1350489	哌唑嗪	NO	YES	NO
1351557	坎地沙坦	NO	YES	NO
1353766	普萘洛尔	NO	YES	NO
1353776	非洛地平	NO	YES	NO
1363053	多沙唑嗪	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1367500	氯沙坦	NO	YES	NO
1373225	培哚普利	NO	YES	NO
1373928	胼酞嗪	NO	YES	NO
1386957	拉贝洛尔	NO	YES	NO
1395058	氯噻酮	NO	YES	NO
1398937	可乐定	NO	YES	NO

40226742	奥美沙坦	NO	YES	NO
40235485	阿齐沙坦	NO	YES	NO

表 B.14: 高血压疾病

概念 ID	概念名称	排除	后代	映射
316866	高血压疾病	NO	YES	NO

B.7 随访时间大于 3 年的高血压一线治疗患者

与队列定义 2.6 相同，但需在事件索引日期前至少 365 天和之后 1095 天连续观察

B.8 血管紧张素转化酶抑制剂的使用

初始事件队列

有下列任何一种情况的人:

ACE 抑制剂药物暴露(表 B.15)，

在事件索引日期之前至少 0 天和之后 0 天连续观察,并限制初始事件:每个人的所有事件。

限制合格队列:所有事件每个人。

结束日期策略

该策略根据在特定概念集中发现的代码创建一个药物时期。如果在一个时期内发现了索引事件，队列结束日期将使用该时期的结束日期。否则，它将使用包含索引事件的观察期结束日期。

使用 ACE 抑制剂的结束日期(表 B.15)

•曝光间隔为 30 天

•曝光结束后增加 0 天

队列撤销策略

按时间撤销队列，间隔大小为 30 天。

概念集定义

表 B.15: 血管紧张素转换酶抑制剂

概念 ID	概念名称	排除	后代	映射
1308216	利诺普利	NO	YES	NO
1310756	莫西普利	NO	YES	NO
1331235	奎那普利	NO	YES	NO
1334456	雷米普利	NO	YES	NO
1335471	贝那普利	NO	YES	NO
1340128	卡托普利	NO	YES	NO
1341927	依那普利	NO	YES	NO
1342439	曲多普利	NO	YES	NO
1363749	福辛普利	NO	YES	NO
1373225	培哚普利	NO	YES	NO

B.9 血管紧张素受体阻滞剂的使用

与队列定义 B.8 相同，使用血管紧张素受体阻滞剂(ARBs)(表 B.16)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.16: 血管紧张素受体阻滞剂

概念 ID	概念名称	排除	后代	映射
1308842	缬沙坦	NO	YES	NO
1317640	替米沙坦	NO	YES	NO
1346686	依普沙坦	NO	YES	NO
1347384	厄贝沙坦	NO	YES	NO
1351557	坎地沙坦	NO	YES	NO
1367500	氯沙坦	NO	YES	NO
40226742	奥美沙坦	NO	YES	NO
40235485	阿齐沙坦	NO	YES	NO

B.10 噻嗪类或噻嗪类利尿剂的使用

与队列定义 2B.8 相同，使用噻嗪类或噻嗪类利尿剂(表 B.17)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.17: 噻嗪类或噻嗪类利尿剂

概念 ID	概念名称	排除	后代	映射
907013	美托拉宗	NO	YES	NO
974166	氢氯噻嗪	NO	YES	NO
978555	吲达帕胺	NO	YES	NO
1395058	氯噻酮	NO	YES	NO

B.11 二氢吡啶钙通道阻滞剂(dCCB)的使用

与队列定义 B.8 相同，用二氢吡啶钙通道阻滞剂(dCCB) (表 B.18)代替 ACE 抑制剂(表 B.15)

概念集定义

表 B.18: 二氢吡啶钙通道阻滞剂

概念 ID	概念名称	排除	后代	映射
1318137	尼卡地平	NO	YES	NO
1318853	硝苯地平	NO	YES	NO
1319880	尼索地平	NO	YES	NO
1326012	伊拉地平	NO	YES	NO
1332418	氨氯地平	NO	YES	NO
1353776	非洛地平	NO	YES	NO

B.12 非二氢吡啶钙通道阻滞剂(nd- CCB)的使用

与队列定义 B.8 相同，使用非二氢吡啶钙通道阻滞剂(ndCCB)(表 B.19)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.19: 非二氢吡啶钙通道阻滞剂

概念 ID	概念名称	排除	后代	映射
1307863	维拉帕米	NO	YES	NO
1328165	地尔硫卓	NO	YES	NO

B.13 β受体阻滞剂的使用

与队列定义 B.8 相同，使用β受体阻滞药(表 B.20)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.20: β受体阻滞剂

概念 ID	概念名称	排除	后代	映射
1307046	美托洛尔	NO	YES	NO
1313200	纳多洛尔	NO	YES	NO
1314002	阿替洛尔	NO	YES	NO
1314577	奈比洛尔	NO	YES	NO
1319998	醋丁洛尔	NO	YES	NO
1322081	倍他洛尔	NO	YES	NO
1327978	喷布洛尔	NO	YES	NO
1338005	比索洛尔	NO	YES	NO
1345858	吲哚洛尔	NO	YES	NO
1346823	卡维地洛	NO	YES	NO
1353766	普萘洛尔	NO	YES	NO
1386957	拉贝洛尔	NO	YES	NO

B.14 袢利尿剂的使用

与队列定义 B.8 相同，使用袢利尿剂(表 B.21)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.21: 袢利尿剂

概念 ID	概念名称	排除	后代	映射
932745	布美他尼	NO	YES	NO
942350	托拉塞米	NO	YES	NO
956874	呋塞米	NO	YES	NO

B.15 保钾利尿剂的使用

与队列定义 B.8 相同，使用保钾利尿剂(表 B.22)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.22: 保钾利尿剂

概念 ID	概念名称	排除	后代	映射
904542	氨苯蝶啶	NO	YES	NO

991382	阿米洛利	NO	YES	NO

B.16 α-1 阻断剂的使用

与队列定义 B.8 相同，使用α-1 阻滞剂(表 B.23)代替 ACE 抑制剂(表 B.15)。

概念集定义

表 B.23: α-1 阻滞剂

概念 ID	概念名称	排除	后代	映射
1341328	特拉唑嗪	NO	YES	NO
1350489	哌唑嗪	NO	YES	NO
1363053	多沙唑嗪	NO	YES	NO

C 阴性对照

本附录包含本书各章节中使用的阴性对照。

表 C.1: ACE 抑制剂(ACEi)与噻嗪类和类噻嗪类利尿剂比较的阴性对照结果

概念 ID	名称
434165	Abnormal cervical smear 宫颈刮片异常
436409	Abnormal pupil 瞳孔异常
199192	Abrasion and/or friction burn of trunk without infection 干擦伤或摩擦烧伤，无感染
4088290	Absence of breast 乳房缺如
4092879	Absent kidney 肾脏缺如
44783954	Acid reflux 胃酸倒流
75911	Acquired hallux valgus 拇指感染外翻
137951	Acquired keratoderma 皮肤感染角化病
77965	Acquired trigger finger 触发手指感染
376707	Acute conjunctivitis 急性结膜炎
4103640	Amputated foot 脚部截肢
73241	Anal and rectal polyp 肛门和直肠息肉
133655	Burn of forearm 前臂烧伤
73560	Calcaneal spur 跟骨刺痛
434327	Cannabis abuse 大麻滥用
4213540	Cervical somatic dysfunction 宫颈躯体功能障碍
140842	Changes in skin texture 皮肤纹理变化
81378	Chondromalacia of patella 髌骨软骨软化症
432303	Cocaine abuse 可卡因滥用
4201390	Colostomy present 结肠造萎

46269889	Complication due to Crohn’s disease 克罗恩氏病并发症
概念 ID	名称
134438	Contact dermatitis 接触性皮炎
78619	Contusion of knee 膝盖撞伤
201606	Crohn’s disease 克罗恩氏病
76786	Derangement of knee 膝盖错位
4115402	Difficulty sleeping 失眠
45757370	Disproportion of reconstructed breast 乳房非比例重建
433111	Effects of hunger 饥饿效应
433527	Endometriosis 子宫内膜异位
4170770	Epidermoid cyst 表皮样囊肿
4092896	Feces contents abnormal 粪便含量异常
259995	Foreign body in orifice 孔内异物
40481632	Ganglion cyst 神经节囊肿
4166231	Genetic predisposition 遗传倾向
433577	Hammer toe 锤状趾
4231770	Hereditary thrombophilia 遗传性易栓症
440329	Herpes zoster without complication 无并发症带状疱疹
4012570	High risk sexual behavior 高危性行为
4012934	Homocystinuria 高胱氨酸尿症
441788	Human papilloma virus infection 人乳头瘤病毒感染
4201717	Ileostomy present 现阶段回肠造瘘
374375	Impacted cerumen 耳垢栓塞
4344500	Impingement syndrome of shoulder region 肩部撞击综合征
139099	Ingrowing nail 指甲内生
444132	Injury of knee 膝盖受伤
196168	Irregular periods 月经不调
432593	Kwashiorkor 严重营养不良
434203	Late effect of contusion 挫伤迟发效应
438329	Late effect of motor vehicle accident 机动车祸迟发效应
195873	Leukorrhea 白带
4083487	Macular drusen 黄斑点
4103703	Melena 黑粪症
4209423	Nicotine dependence 尼古丁依赖症
377572	Noise effects on inner ear 噪声对内耳的影响
40480893	Nonspecific tuberculin test reaction 非特异性结核菌素试验反应
136368	Non-toxic multinodular goiter 无毒多结节性甲状腺肿
140648	Onychomycosis due to dermatophyte 由皮肤真菌引起的甲真菌病

438130 Opioid abuse 阿片类药物滥用

概念 ID 名称

4091513 Passing flatus 暂时性肠胃胀气

4202045 Postviral fatigue syndrome 病毒后疲劳综合征

373478 Presbyopia 老花眼

46286594 Problem related to lifestyle 生活方式相关问题

439790 Psychalgia 精神性疼痛

81634 Ptotic breast 内窥镜胸

380706 Regular astigmatism 规则性散光

141932 Senile hyperkeratosis 高龄导致的角化过度

36713918 Somatic dysfunction of lumbar region 腰椎区躯体功能障碍

443172 Splinter of face, without major open wound 面部碎裂，无大开放性伤口

81151 Sprain of ankle 脚踝扭伤

72748 Strain of rotator cuff capsule 肌腱套囊应变

378427 Tear film insufficiency 泪膜功能不全

437264 Tobacco dependence syndrome 烟草依赖综合征

194083 Vaginitis and vulvovaginitis 阴道炎和外阴炎

140641 Verruca vulgaris 寻常疣

440193 Wrist drop 手垂症

4115367 Wrist joint pain 手腕关节痛

D 研究方案模版

1. Table of contents 目录表

2. List of abbreviations 缩写列表

3. Abstract 摘要

4. Amendments and Updates 修订和更新

5. Milestones 重要事件

6. Rationale and Background 理论基础和背景

7. Study Objectives 学习目标

Primary Hypotheses 主要假设

Secondary Hypotheses 次要假设

Primary Objectives 主要目标

Secondary Objectives 次要目标

8. Research methods 研究方法

Study Design 研究设计

Data Source(s)数据来源

Study population 研究群体

Exposures 暴露情况

Outcomes 结果

Covariates 协变量

9. Data Analysis Plan 数据分析计划

Calculation of time-at risk 风险时间的计算

Model Specification 模式设定

Pooling effect estimates across databases 跨数据库的池效应估计

Analyses to perform 分析执行

Output 输出

Evidence Evaluation 评估证据

10. Study Diagnostics 研究诊断

Sample Size and Study Power 样本量和研究效力

Cohort Comparability 队列可比性

Systematic Error Assessment 系统误差评估

11. Strengths and Limitations of the Research Methods 研究方法的优缺点

12. Protection of Human Subjects 受试人员保护

13. Management and Reporting of Adverse Events and Adverse Reactions 不良事件和不良反应的管理和报告

14. Plans for Disseminating and Communicating Study Results 宣传和交流研究结果的计划

15. Appendix: Negative controls 附录:阴性对照

16. References 参考文献

E 参考答案

本附录包含本书中练习题的参考答案。

E1. 通用数据模型

练习 4.1

根据练习中的描述，John 的记录应类似于表 E.1.

表 E.1: PERSON 表

列名	值	解释
PERSON_ID	2	一个唯一的整型值
GENDER_CONCEPT_ID	8507	男性的概念 ID 是 8507
YEAR_OF_BIRTH	1974	
MONTH_OF_BIRTH	8	
DAY_OF_BIRTH	4	
BIRTH_DATETIME	1974-08-04 00:00:00	当时间未知时，使用午夜时间
DEATH_DATETIME	NULL	
RACE_CONCEPT_ID	8516	黑人或非裔美国人的概念ID为8516
ETHNICITY_ CONCEPT_ID	38003564	38003564 指“不是西班牙裔”

LOCATION_ID		他的地址未知
PROVIDER_ID		他的主要保健提供者未知
CARE_SITE		他的主要护理地点未知
PERSON_SOURCE_ VALUE	NULL	未提供
GENDER_SOURCE_ VALUE	Man	描述时使用的文本
GENDER_SOURCE_ CONCEPT_ID	0	
RACE_SOURCE_ VALUE	African American	描述时使用的文本
RACE_SOURCE_ CONCEPT_ID	0	
ETHNICITY_SOURCE_ VALUE	NULL	
ETHNICITY_SOURCE_ CONCEPT_ID	0	

练习 4.2

根据练习中的描述，John 的记录应类似于表 E.2.

表 E.2: OBSERVATION_PERIOD 表

列名	值	解释
OBSERVATION_ PERIOD_ID	2	一个唯一的整型值。
PERSON_ID	2	这是 John 在 PERSON 表中记录的外键。
OBSERVATION_PERIOD_ START_DATE	2015-01-01	入组日期。
OBSERVATION_PERIOD_ END_DATE	2019-07-01	在出组日期之后，不应该再有数据。
PERIOD_TYPE_ CONCEPT_ID	44814722	44814724 指“参加保险期间”。

练习 4.3

根据练习中的描述，John 的记录应类似于表 E.3.

表 E.3: DRUG_EXPOSURE 表

列名	值	解释
DRUG_EXPOSURE_ID	1001	一个唯一的整型值。
PERSON_ID	2	这是 John 在 PERSON 表中记录的外键。
DRUG_CONCEPT_ID	19078461	提供的 NDC 代码映射到标准概念 19078461。

DRUG_EXPOSURE_ START_DATE	2019-05-01	药物暴露的开始日期。
DRUG_EXPOSURE_ START_DATETIME	2019-05-01 00:00:00	当时间未知时，使用午夜时间。
DRUG_EXPOSURE_ END_DATE	2019-05-31	基于开始日期+天数。
DRUG_EXPOSURE_ END_DATETIME	2019-05-31 00:00:00	当时间未知时，使用午夜时间。
VERBATIM_END_DATE	NULL	未提供
DRUG_TYPE_ CONCEPT_ID	38000177	38000177 表示 “处方已写”。
STOP_REASON	NULL	
REFILLS	NULL	
QUANTITY	NULL	未提供。
DAYS_SUPPLY	30	如练习中所述。
SIG	NULL	未提供。
ROUTE_CONCEPT_ID	4132161	4132161 表示“口服”。
LOT_NUMBER	NULL	未提供。
PROVIDER_ID	NULL	未提供。
VISIT_OCCURRENCE_ ID	NULL	本次就诊未提供信息。
VISIT_DETAIL_ID	NULL	
DRUG_SOURCE_ VALUE	76168009520	这是提供的 NDC 代码。
DRUG_SOURCE_ CONCEPT_ID	583945	583945 代表药物来源值（NDC 代码“76168009520”）。
ROUTE_SOURCE_ VALUE	NULL	

练习 4.4

要查找记录集，我们可以查询 CONDITION_OCCURRENCE 表：

```
library(DatabaseConnector)
connection<-connect(connectionDetails)
sql <- "SELECT *
FROM @cdm.condition_occurrence

WHERE condition_concept_id =192671;"

result <- renderTranslateQuerySql(connection, sql, cdm = "main")
head(result)
```

```
##                    CONDITION_OCCURRENCE_ID PERSON_ID
CONDITION_CONCEPT_ID ...
          ## 1                    4657                273                192671 ...
```

## 2	1021	61	192671 ...
## 3	5978	351	192671 ...
## 4	9798	579	192671 ...
## 5	9301	549	192671 ...
## 6	1997	116	192671 ...

练习 4.5

```
library(DatabaseConnector)
connection                          <-
connect(connectionDetails)   sql    <-
"SELECT *
FROM   @cdm.observation_period   WHERE
person_id = 61;"
renderTranslateQuerySql(connection,
sql, cdm = "main")
```

要查找数据记录集，我们可以用 CONDITION_SOURCE_VALUE 字段查询 CONDITION_OCCURRENCE 表：

CONDITION_OCCURRENCE_ID PERSON_ID CONDITION_CONCEPT_ID ...

```
sql <- "SELECT *
FROM @cdm.condition_occurrence
WHERE    condition_source_value    =
'K92.2';"

result                            <-
renderTranslateQuerySql(connection,
sql, cdm = "main")
head(result)
```

## 1	4657	273	192671	...
## 2	1021	61	192671	...
## 3	5978	351	192671	...
## 4	9798	579	192671	...
## 5	9301	549	192671	...
## 6	1997	116	192671	...

练习 4.6

这个信息保存在 OBSERVATION_PERIOD 表中：

```
##   OBSERVATION_PERIOD_ID PERSON_ID OBSERVATION_PERIOD_START_DATE ...
## 1                    61        61                   1968-01-21 ...
```

E2. 标准化词汇

练习 5.1

概念 ID 192671（“胃肠道出血”）

练习 5.2

ICD-10CM 代码:

- K29.91 “未明确的胃十二指肠炎，有出血”
- K92.2 “胃肠道出血，未特指”

ICD-9CM 代码:

- 578 “胃肠道出血”
- 578.9 “未明确的胃肠道出血”

练习 5.3

MedDRA 首选词:

- “胃肠道出血” (概念 ID 35707864)
- “肠出血” (概念 ID 35707858)

E3. 抽取、转换、加载

练习 6.1

A) 数据专家和 CDM 专家共同设计 ETL
B) 有医学知识的人创建代码映射
C) 技术人员实施 ETL
D) 所有人都参与质量控制

练习 6.2

列名	值	解释
PERSON_ID	A123B456	该列的数据类型为整数，因此源记录值需要转换为数字值。
GENDER_CONCEPT_ID	8532	
YEAR_OF_BIRTH	NULL	如果出生的月份或日期未知，不要猜测。数据中允许没有出生月份或日期的人存在。如果一个人没有出生年份，则该人应该被舍弃。由于没有出生年份，她也将不得不被从分析中舍弃。
MONTH_OF_BIRTH	NULL	
DAY_OF_BIRTH	NULL	
RACE_CONCEPT_ID	0	种族是白种人，应该被映射到 8527。
ETHNICITY_CONCEPT_ ID	8527	未提供种族信息，应将其映射为 0。
PERSON_SOURCE_ VALUE	A123B456	
GENDER_SOURCE_ VALUE	F	

RACE_SOURCE_VALUE	WHITE
ETHNICITY_SOURCE_VALUE	NONE PROVIDED

练习 6.3

列名	值
VISIT_OCCURRENCE_ID	1
PERSON_ID	11
VISIT_START_DATE	2004-09-26
VISIT_END_DATE	2004-09-30
VISIT_CONCEPT_ID	9201
VISIT_SOURCE_VALUE	inpatient

E4. 数据分析用例

练习 7.1

特征刻画

患者水平的预测

人群水平的估计

练习 7.2

可能不会。由于人们服用双氯芬酸通常是有原因的， 定义一组与服用双氯芬酸人群相似的未使用过双氯芬酸的人群通常是不可能的。这限制了人与人之间的比较。个体比较也许是可能的。对于服用双氯芬酸人群中的每个患者，可以找出他们没有服用双氯芬酸的一段时间进行比较。但在这里也会出现类似的问题：这些时间可能是无法比较的，因为某些原因，一个人在某一时间会服用双氯芬酸而其他时候则不会。

E5. SQL 与 R

练习 9.1

要计算人数，我们可以简单地查询 PERSON 表:

```
library(DatabaseConnector)
connection                    <-
connect(connectionDetails) sql <-
"SELECT COUNT(*) AS person_count
FROM @cdm.person;"

renderTranslateQuerySql(connectio
n, sql, cdm = "main")
```

```
##
PERSON_COUNT
## 1  2694
```

练习 9.2

要计算至少收到一次塞来昔布处方的人数，我们可以查询 DRUG_EXPOSURE 表。要找到所有含有塞来昔布成分的药物，我们关联了 CONCEPT_ANCESTOR 和 CONCEPT 表:

```
library(DatabaseConnector)
connection                          <-
connect(connectionDetails)
sql             <-             "SELECT
COUNT(DISTINCT(person_id))         AS
person_count FROM @cdm.drug_exposure
INNER JOIN @cdm.concept_ancestor
ON          drug_concept_id          =
descendant_concept_id    INNER    JOIN
@cdm.concept ingredient
ON        ancestor_concept_id        =
ingredient.concept_id           WHERE
LOWER(ingredient.concept_name)      =
'celecoxib'                        AND
ingredient.concept_class_id         =
'Ingredient'
AND   ingredient.standard_concept   =
'S';"

renderTranslateQuerySql(connection, sql, cdm ="main")
```

##

PERSON_COUNT

1 1844

请注意，考虑到一个人可能有多个处方，我们使用 COUNT(DISTINCT(person_id))来查找不同的人的数量。另请注意，我们使用 LOWER 函数使搜索“celecoxib”不区分大小写。

或者，我们可以使用 DRUG_ERA 表，该表已经汇总到成分级别：

```
library(DatabaseConnector)
connection                          <-
connect(connectionDetails)

sql             <-             "SELECT
COUNT(DISTINCT(person_id))         AS
person_count FROM @cdm.drug_era
INNER JOIN @cdm.concept ingredient
ON          drug_concept_id          =
ingredient.concept_id           WHERE
LOWER(ingredient.concept_name)      =
'celecoxib'
AND   ingredient.concept_class_id   =
'Ingredient'                      AND
ingredient.standard_concept = 'S';"

renderTranslateQuerySql(connection, sql, cdm = "main")
```

##

PERSON_COUNT

1 1844

练习 9.3

为了计算暴露期间的诊断数量，我们通过关联到 CONDITION_OCCURRENCE 表来扩展先前的查询。我们关联到 CONCEPT_ANCESTOR 表以查找所有提示胃肠道出血的病情概念:

```
library(DatabaseConnector)
connection                      <-
connect(connectionDetails) sql <-
"SELECT        COUNT(*)        AS
diagnose_count FROM @cdm.drug_era
INNER      JOIN      @cdm.concept
ingredient
ON        drug_concept_id        =
ingredient.concept_id  INNER  JOIN
@cdm.condition_occurrence
ON     condition_start_date     >=
drug_era_start_date           AND
condition_start_date           <=
drug_era_end_date
INNER JOIN @cdm.concept_ancestor
ON           condition_concept_id
=descendant_concept_id      WHERE
LOWER(ingredient.concept_name)  =
'celecoxib'                   AND
ingredient.concept_class_id     =
'Ingredient'
AND  ingredient.standard_concept =
'S' ANDancestor_concept_id=192671;"

renderTranslateQuerySql(connection, sql, cdm ="main")
```

##

DIAGNOSE_COUNT

1 41

请注意，在这种情况下，必须使用 DRUG_ERA 表而不是 DRUG_EXPOSURE 表，因为具有相同成分的药物暴露可以重叠，但是药物成分时段不能重叠。这可能导致重复计算。例如，假设一个人同时接受了两种含有塞来昔布的药物。这将被记录为两次药物暴露，因此在暴露期间发生的任何诊断都将被计数两次。两次暴露将合并为一个不重叠的药物成分时段。

E6. 定义队列

练习 10.1

我们创建满足以下这些要求的初始事件条件：

- 第一次使用双氯芬酸
- 16 岁或以上
- 暴露前至少 365 天连续观察

完成后，进入事件部分的队列应如图 E.1 所示。

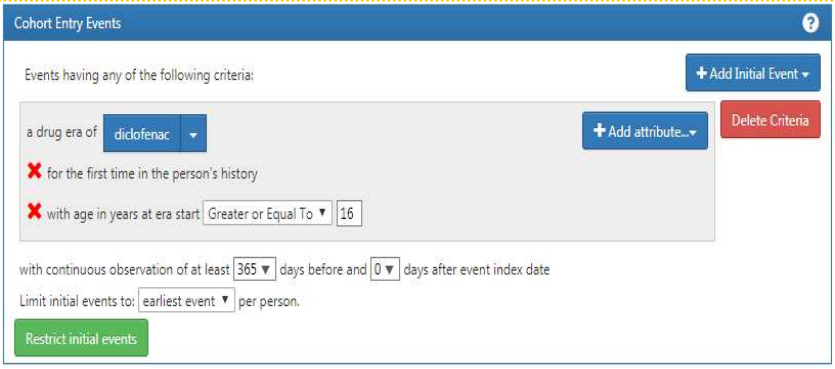


图 E.1：双氯芬酸新用户的队列入组事件设置

双氯芬酸的概念集表达应如图 E.2 所示，包括成分“双氯芬酸”及其所有衍生概念，因此应包括所有含有双氯芬酸成分的药物。

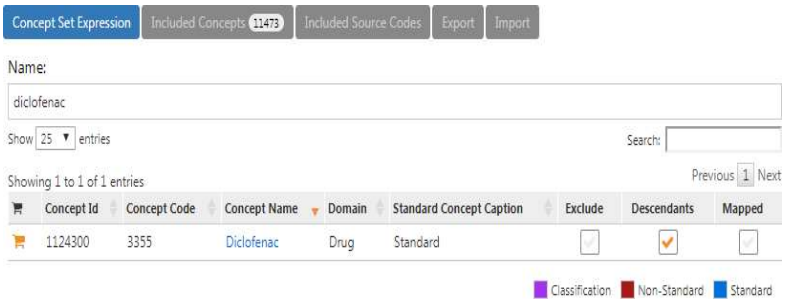


图 E.2：双氯芬酸的概念集表达

接下来，我们要求先前不能接触任何 NSAID，如图 E.3 所示。

图 E.3：要求先前不能接触任何 NSAID

NSAID 的概念集表达应类似于图 E.4，包括 NSAID 类及其所有衍生概念，因此包括所有包含任何一种 NSAID 的药物。

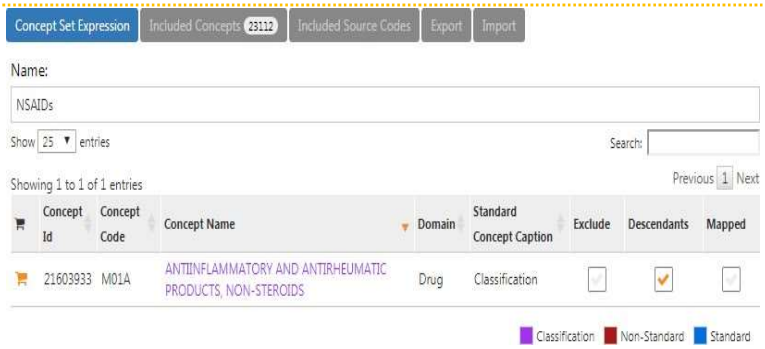


图 E.4: NSAID 的概念集表达式

此外，我们要求先前不能诊断出癌症，如图 E.5 所示。

图 E.5：要求先前不能有癌症诊断

“广泛恶性肿瘤”的概念集表达应如图 E.6 所示，其中包括上层概念“恶性肿瘤疾病”及其所有衍生概念。

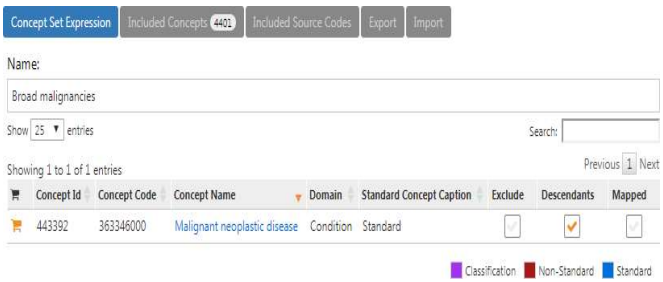


图 E.6：广泛恶性肿瘤的概念集表达

最后，我们将队列退出条件定义为停止暴露（允许 30 天的间隔），如图 E.7 所示。

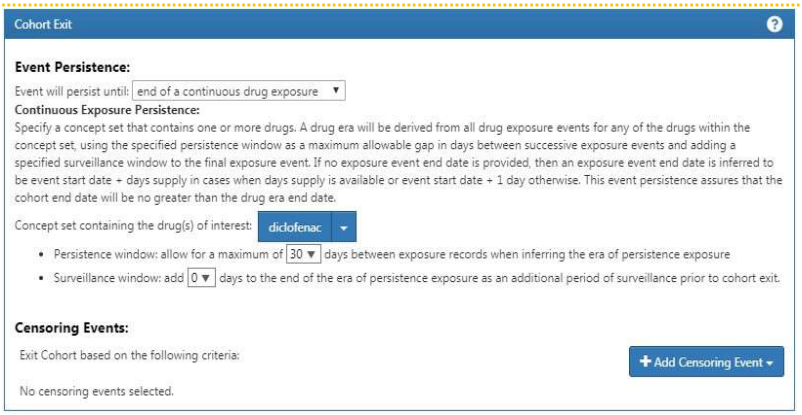


图 E.7：设置队列退出日期

练习 10.2

为了便于阅读，我们在这里将 SQL 分为两个步骤。我们首先找到所有发生心肌梗塞的情况，并将它们存储在名为“ #diagnoses”的临时表中：

```
library(DatabaseConnector
)
connection              <-
connect(connectionDetails
) sql <- "SELECT person_id
AS subject_id,
condition_start_date   AS
cohort_start_date    INTO
#diagnoses
FROM
@cdm.condition_occurrence
WHERE
condition_concept_id IN (
SELECT
descendant_concept_id
FROM
@cdm.concept_ancestor
WHERE ancestor_concept_id
= 4329847 -- Myocardial
infarction
)
AND condition_concept_id
NOT    IN    (    SELECT
descendant_concept_id
FROM
@cdm.concept_ancestor
WHERE ancestor_concept_id
= 314666 -- Old myocardial
infarction
);"

renderTranslateExecuteSql
```

然后，我们仅选择住院或急诊就诊时发生的情况，并且使用某种唯一的 COHORT_DEFINITION_ID（我们选择为“ 1”）：

```
sql    <-    "INSERT    INTO
@cdm.cohort ( subject_id,
cohort_start_date,
cohort_definition_id
)
SELECT           subject_id,
cohort_start_date,
CAST    (1    AS    INT)    AS
cohort_definition_id    FROM
#diagnoses
INNER                    JOIN
@cdm.visit_occurrence     ON
subject_id = person_id
AND   cohort_start_date   >=
visit_start_date         AND
cohort_start_date         <=
visit_end_date
WHERE   visit_concept_id   IN
(9201,    9203,    262);    --
Inpatient or ER;"

renderTranslateExecuteSql(co
nnection, sql, cdm = "main")
```

请注意，一种替代方法是根据 VISIT_OCCURRENCE_ID 将条件与就诊进行关联，而不是要求条件日期在就诊开始和结束日期之内。这可能会更准确，因为它将保证所记录的情况是与住院或急诊相关的。但是，许多观察性数据库没有记录就诊和诊断之间的关联关系，因此我们选择使用日期来代替。这可能会给我们带来更高的敏感性，但特异性可能会低一些。

另请注意，我们忽略了队列的结束日期。通常，当使用队列定义结果时，我们只对队列开始日期感兴趣，而创建（未定义的）队列结束日期毫无意义。

建议在不再需要时清理所有临时表：

E7. 特征刻画

```
sql   <-   "TRUNCATE   TABLE
#diagnoses;    DROP    TABLE
#diagnoses;"

renderTranslateExecuteSql(
connection, sql)
```

练习 11.1

在 ATLAS 中，我们点击 

并选择感兴趣的数据源。我们可以选择药物暴露报告，选择“Table”标签，然后搜索“塞来昔布”，如图 E.8 所示。在这里，我们看到这个特定的数据库展示了塞来昔布的各种制剂。我们可以单击任何一种药物以获得更详细的视图，例如显示这些药物的年龄和性别分布。

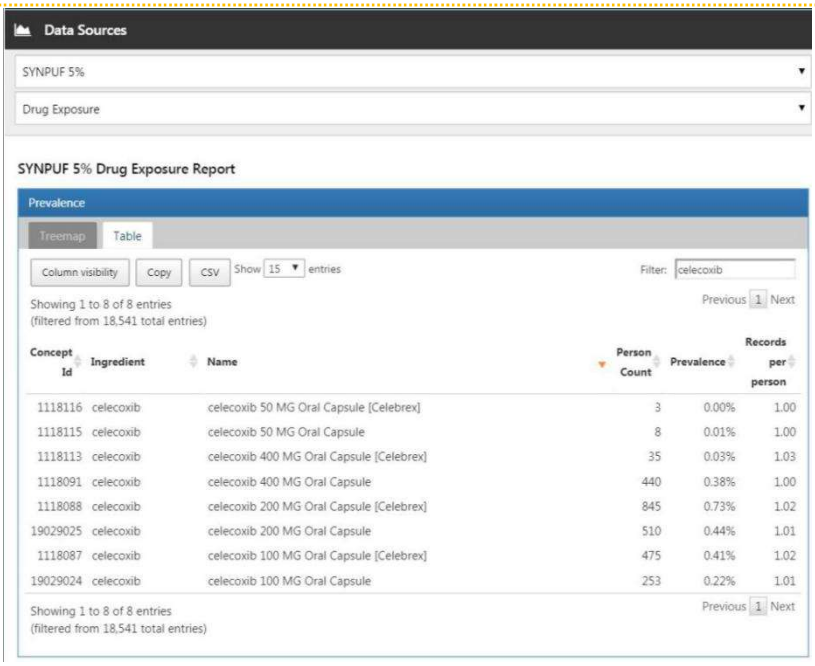


图 E.8：数据源特征刻画

练习 11.2

点击 Cohort Definitions ，然后点击“New cohort”来创建一个新的队列。给队列取一个有意义的名称（例如“celecoxib 新用户”），然后转到“概念集”标签。点击“New Concept Set”，为概念集起一个有意义的名称（例如“celecoxib”）。打开 Search 模块，搜索“celecoxib”，将级别限制为“Ingredient”，将标准概念 限制为“Standard”，然后单击 ，将概念添加到您的概念集中，如图 E.9 所示。

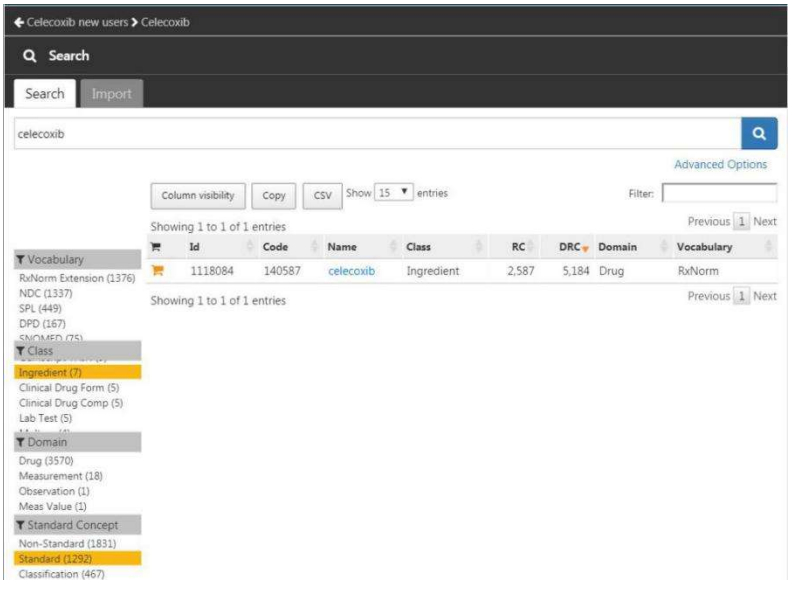


图 E.9：选择“塞来昔布”成分的标准概念

点击图 E.9 左上方所示的左箭头，以返回您的队列定义。点击“+Add Initial Event”，然后点击“Add Drug Era”，选择您先前为药品成分时段标准创建的概念集。点击“Add attribute…”，然后选择“Add First Exposure Criteria”，将所需的连续观察设置为索引日期之前至少 365 天。结果应类似于图 E.10。“Inclusion Criteria”、“Cohort Exit”和“Cohort Eras”部分保持原来的样子。确保通过点击  来保存队列定义，然后通过点击  将其关闭。

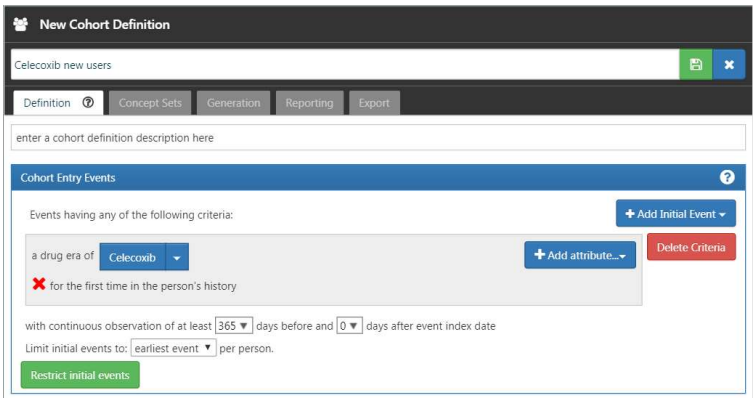
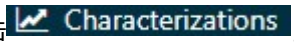


图 E.10：一个简单的塞来昔布新用户队列定义

现在我们已经定义了队列，就可以对其进行表征了。点击 

，然后单击“New Characterization”。为您的特征刻画起一个有意义的名称（例如“Celecoxib 新用户特征刻画”）。在“Cohort Definitions”下，点击“Import”，然后选择您最近创建的队列定义。在“Feature Analyses”下，单击“Import”，并至少选择一种条件分析和一种药物分析，例如“Drug Group Era Any Time Prior”和“Condition Group Era Any Time Prior”。现在，您的特征刻画定义应类似于图 E.11。确保通过单击  来保存特征刻画的设置。

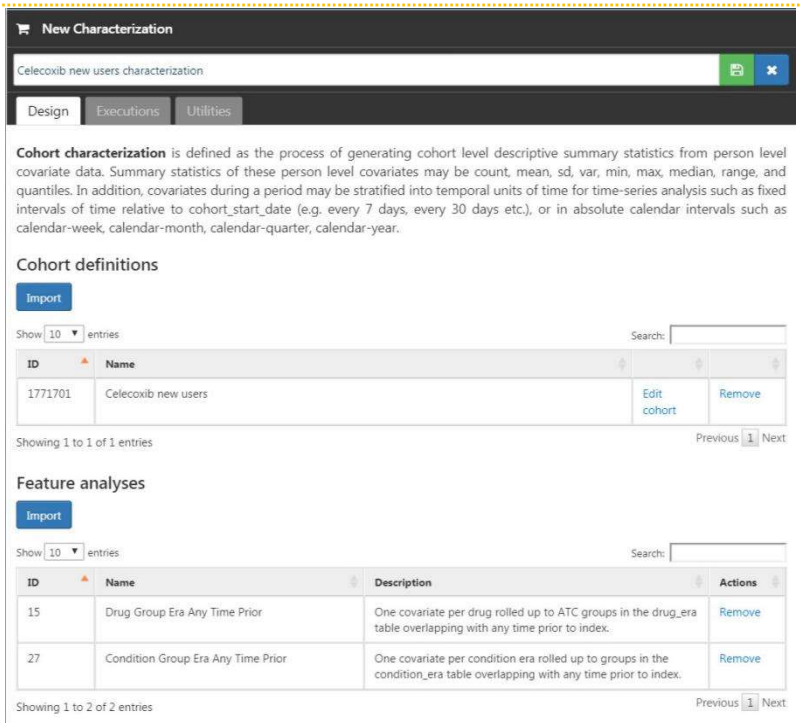


图 E.11：特征刻画的设置

点击“Exections”选项卡，然后点击“Generate”以获取数据源。数据可能需要一段时间才能生成。完成后，我们可以点击“View latest results”。生成的结果将类似于图 E.12，数据显示，像疼痛和关节病通常是可以被看到的，这并不奇怪，因为这些是塞来昔布的适应症。在列表的下方，我们可能会看到一些之前没有想到的情况。

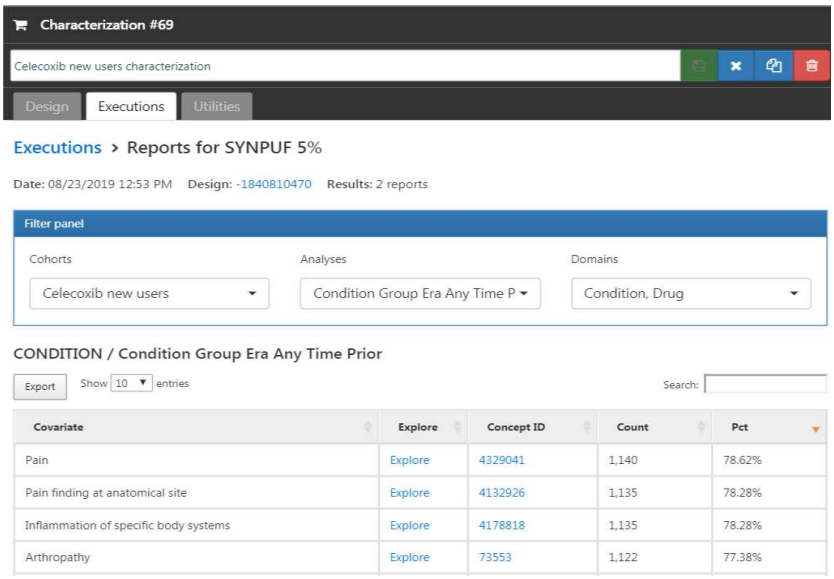


图 E.12：特征刻画的设置

练习 11.3

点击Cohort Definitions，然后点击“New cohort”来创建一个新的队列。给这个队列起一个有意义的名字(如“GI bleed”)，然后跳转到“Concept Sets”选项卡。点击“New Concept Set”，给概念集其一个有意义的名字(如“GI bleed”)。打开Search模块， 搜索“Gastrointestinal hemorrhage”，然后点击顶部概念旁边的按钮，将概念添加到概念集中，如图 E.13 所示。

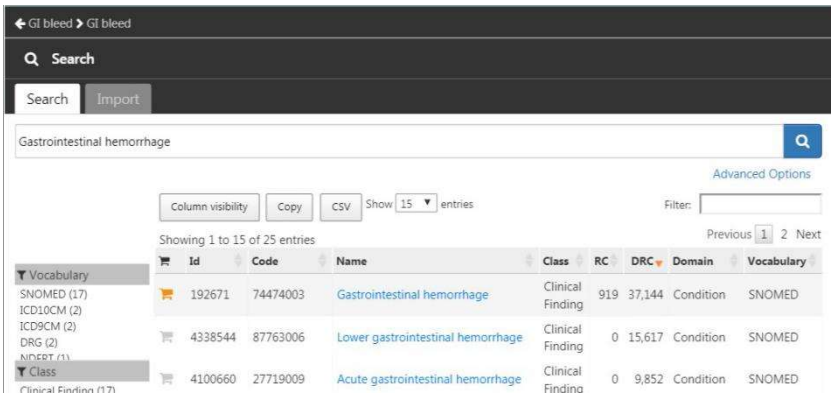


图 E.13：选择“胃肠道出血”的标准概念

点击图 E.13 左上方所示的左箭头，返回到队列的定义。再次打开“Concept Sets”选项卡，然后检查 GI hemorrhage 这个概念旁边的“Descendants”，如图 E.14 所示。

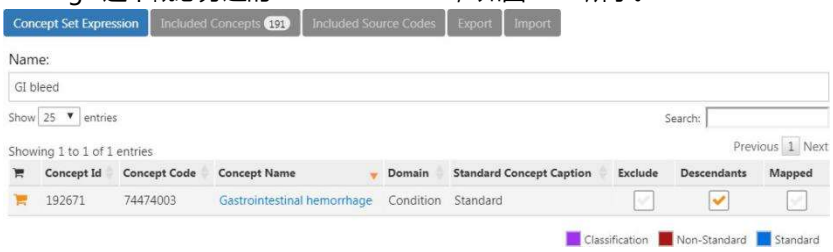


图 E.14：将所有衍生概念添加到“胃肠道出血”中

返回“Definition”选项卡，点击“+Add Initial Event”，然后点击“Add Condition Occurrence”。选择先前创建的用于条件发生条件的概念集。结果应类似于图 E.15。“Inclusion Criteria”、“Cohort Exit”、“Cohort Eras”这几部分保持不动。确保通过单击保存队列定义，然后通过单击将其关闭。

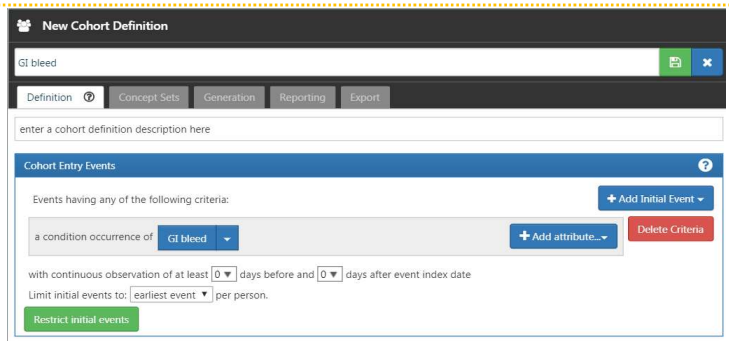


图 E.15：一个简单的胃肠道出血队列的定义

现在我们已经定义了队列，就可以计算发病率了。点击 Incidence Rates，然后点击“New Analysis”，给你的分析其一个有意义的名字(如“Incidence of GI bleed after celecoxib initiation”)。点击“Add Target Cohort”，然后选择我们的 celecoxib 新用户队列。单击“Add Outcome Cohort”，然后添加我们的新 GI 出血队列。将风险时间设置为在开始日期后 1095 天结束。现在，分析应类似于图 E.16。确保通过点击 来保存分析的设置。

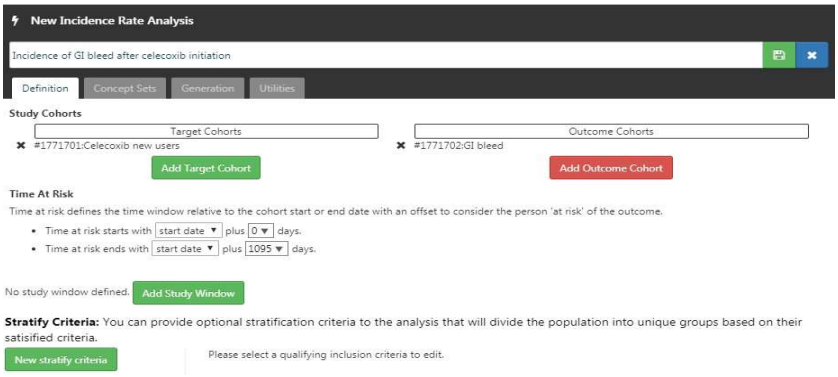


图 E.16：发生率分析

点击“Generation”选项卡，然后点击“Generate”，选择一个数据源并点击“Generate”。 完成后，我们可以看到计算出的发病率和比例，如图 E.17 所示。

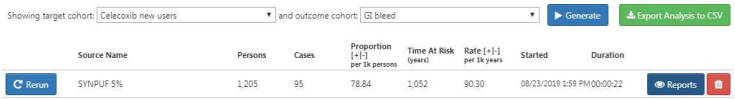


图 E.17：发生率结果

E8. 群体水平评估

练习 12.1

我们指定了默认的一组协变量，但是必须排除正在比较的两种药物（包括它们的所有衍生概念），否则我们的倾向模型将具有完美的预测性：

```
library(CohortMethod)
nsaids <- c(1118084, 1124300) # celecoxib, diclofenac
covSettings                 <-
  createDefaultCovariateSettin
  gs( excludedCovariateConc
  eptIds      =      nsaids,
  addDescendantsToExclude =
  TRUE)
# Load data:
cmData              <-
  getDbCohortMethodDa
  ta( connectionDetail
  s                   =
  connectionDetails,
  cdmDatabaseSchema =
  "main",

  targetId = 1,
  comparatorId = 2,
  outcomeIds =
  3,
  exposureDatab
  aseSchema    =
  "main",
  exposureTabl
  e = "cohort",
  outcomeDataba
  seSchema     =
  "main",
  outcomeTable
  =   "cohort",
  covariateSett
  ings         =
  covSettings)
summary(cmData)
```

```
## CohortMethodData object summary
```

```
   getDbCohortMethodData( connectionDetails = connectionDetails,
cdmDatabaseSchema = "main",
   targetId = 1,
   comparatorId = 2,
   outcomeIds = 3, exposureDatabaseSchema = "main", exposureTable
=  "cohort",  outcomeDatabaseSchema  =  "main",  outcomeTable  =
"cohort", covariateSettings = covSettings)
   summary(cmData)
```

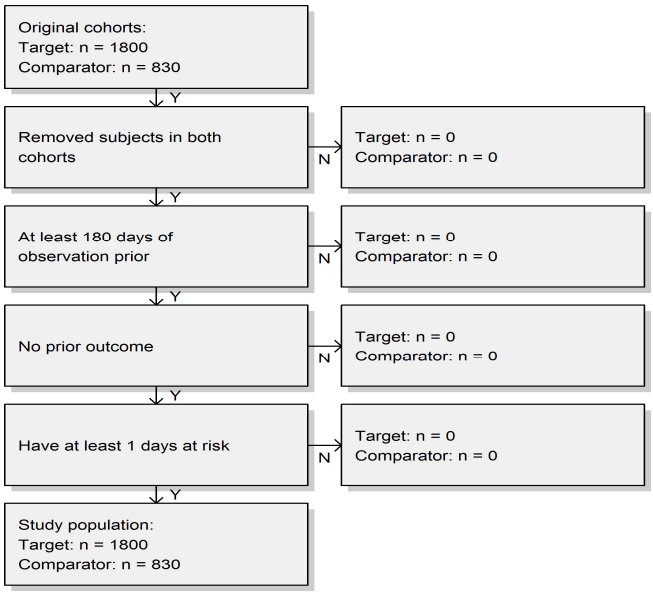
```
##
## Treatment concept ID: 1
## Comparator concept ID: 2
## Outcome concept ID(s): 3
##
## Treated persons: 1800
## Comparator persons: 830
##
## Outcome counts:
```

```
##   Event count Person count
## 3     479          479
##
## Covariates:
## Number of covariates: 389
## Number of non-zero covariate values: 26923
```

练习 12.2

我们按照规范创建研究人群，并输出损耗图：

```
studyPop             <-
  createStudyPopulatio
  n( cohortMethodData =
  cmData,  outcomeId  =
  3,
washoutPeriod   =   180,
  removeDuplicateSubje
  cts  =  "remove  all",
  removeSubjectsWithPr
  iorOutcome   =   TRUE,
  riskWindowStart = 0,
startAnchor   =   "cohort
  start", riskWindowEnd
  = 99999)
drawAttritionDiagram(studyPop)
```



可以看出，与原始队列相比，我们没有丢失任何个体，这可能是因为此处使用的限制条件已在队列定义中应用。

练习 12.3

我们使用 Cox 回归拟合简单的结局模型：

```
model <- fitOutcomeModel(population = studyPop,
                         modelType = "cox")
model
```

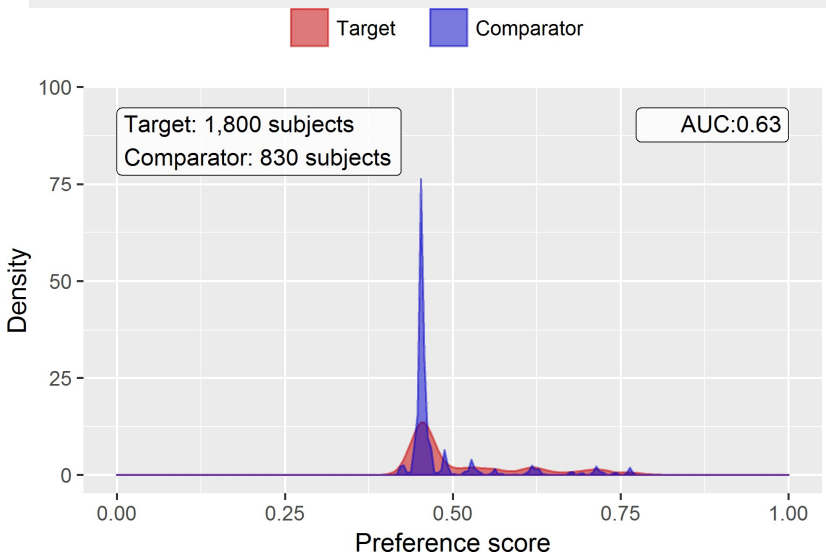
```
## Model type: cox
## Stratified: FALSE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE ## Status: OK
##
##    Estimate lower .95 upper .95  logRr seLogRr
## treatment      1.34612   1.10065   1.65741 0.29723     0.1044
```

塞来昔布使用者可能无法与双氯芬酸使用者互换，并且这些基线差异已经导致不同的结局风险。如果我们不像在此分析中那样针对这些差异进行调整，就有可能产生有偏差的估计。

练习 12.4

我们使用提取的所有协变量，在研究人群中拟合倾向模型，并显示偏好分数分布：

```
ps <- createPs(cohortMethodData = cmData,
population = studyPop)
plotPs(ps, showCountsLabel = TRUE, showAucLabel = TRUE)
```



请注意，这种分布看起来有些奇怪，有几个尖峰。这是因为我们使用的是非常小的模拟数据集。真实的偏好分数分布趋于平滑得多。

倾向模型的 AUC 为 0.63，表明目标人群和比较人群之间存在差异。我们看到两组之间有很多重叠，这表明 PS 调整可以使它们更具可比性。

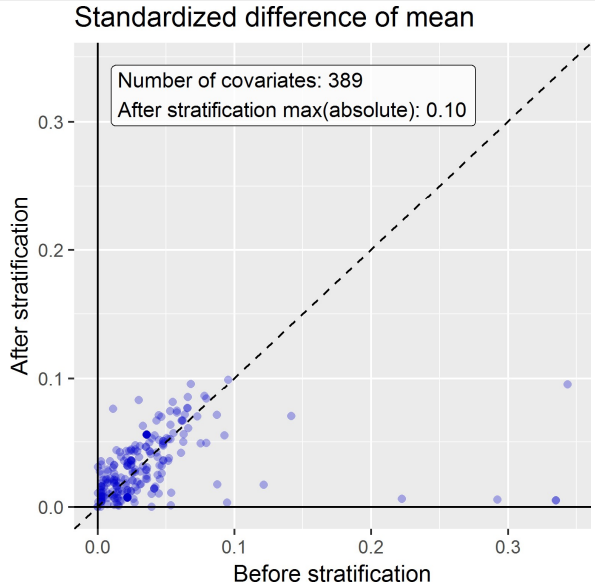
练习 12.5

我们根据倾向性评分对人群进行分层，然后计算分层前后的协变量平衡：

```
strataPop <- stratifyByPs(ps, numberOfStrata = 5) bal <-
computeCovariateBalance(strataPop, cmData)
plotCovariateBalanceScatterPlot(bal,

showCovariateCountLabel = TRUE, showMaxLabel = TRUE,

beforeLabel = "Before stratification", afterLabel = "After
stratification")
```



可以看到，各种基线协变量在分层（x 轴）之前显示出较大的（> 0.3）均值标准差。分层后，平衡性增加了，最大标准差<= 0.1。

练习 12.6

我们使用 Cox 回归拟合结果模型，但通过 PS 分层对其进行分层：

```
adjModel <- fitOutcomeModel(population = strataPop,
modelType = "cox", stratified = TRUE)
adjModel
```

```
## Model type: cox
## Stratified: TRUE
## Use covariates: FALSE
## Use inverse probability of treatment weighting: FALSE
## Status: OK
##
```

```
##     Estimate lower .95 upper .95  logRr seLogRr
## treatment      1.13211    0.92132    1.40008 0.12409     0.1068
```

可以看到，调整后的估计值低于未调整的估计值，并且 95%的置信区间现在包括 1。这是因为我们调整了两个暴露组之间的基线差异，从而减少了偏倚。

E9. 患者水平预测

练习 13.1

我们指定一组协变量设置，并使用 getPlpData 函数从数据库中提取数据：

```
library(PatientLevelPrediction)
covSettings <- createCovariateSettings(
  useDemographicsGender = TRUE,
  useDemographicsAge = TRUE,
  useConditionGroupEraLongTerm = TRUE,
  useConditionGroupEraAnyTimePrior = TRUE,
  useDrugGroupEraLongTerm = TRUE,
  useDrugGroupEraAnyTimePrior = TRUE,
  useVisitConceptCountLongTerm = TRUE,
  longTermStartDays = -365,
  endDays = -1)
```

plpData object summary

```
plpData <- getPlpData(connectionDetails = connectionDetails,
                      cdmDatabaseSchema = "main",
                      cohortDatabaseSchema = "main",
                      cohortTable = "cohort",
                      cohortId = 4,
                      covariateSettings = covSettings,
                      outcomeDatabaseSchema = "main",
                      outcomeTable = "cohort",
                      outcomeIds = 3)

summary(plpData)
```

##

At risk cohort concept ID: -1 ## Outcome concept ID(s):

##

People: 2630

##

```
## Outcome counts:
##    Event count Person count
## 3          479          479
##
## Covariates:
## Number of covariates: 245
## Number of non-zero covariate values: 54079
```

练习 13.2

我们为感兴趣的结局创建了一个研究人群（在这种情况下，这是我们提取数据的唯一结局），除去在开始 NSAID 之前经历了结局的人，并且需要 364 天的风险时间：

```
population <- createStudyPopulation(plpData = plpData,
outcomeId = 3,
washoutPeriod = 364, firstExposureOnly = FALSE,
removeSubjectsWithPriorOutcome = TRUE, priorOutcomeLookback = 9999,
riskWindowStart = 1,
riskWindowEnd = 365, addExposureDaysToStart = FALSE,
addExposureDaysToEnd = FALSE, minTimeAtRisk = 364, requireTimeAtRisk
= TRUE, includeAllOutcomes =TRUE)
nrow(population)
```

```
##
[1] 2578
```

在这种情况下，通过删除先前有结局的人，并且要求至少有 364 天的风险时间，我们舍弃了一些人。

练习 13.3

我们通过首先创建模型设置对象，然后调用 runPlp 函数来运行 LASSO 模型。在这种情况下，我们进行人员拆分，在 75%的数据上训练模型并在 25%的数据上进行评估：

```
lassoModel <- setLassoLogisticRegression(seed = 0)

lassoResults <- runPlp(population = population,
plpData = plpData, modelSettings = lassoModel, testSplit = 'person',
testFraction = 0.25,
nfold = 2,
splitSeed = 0)
```

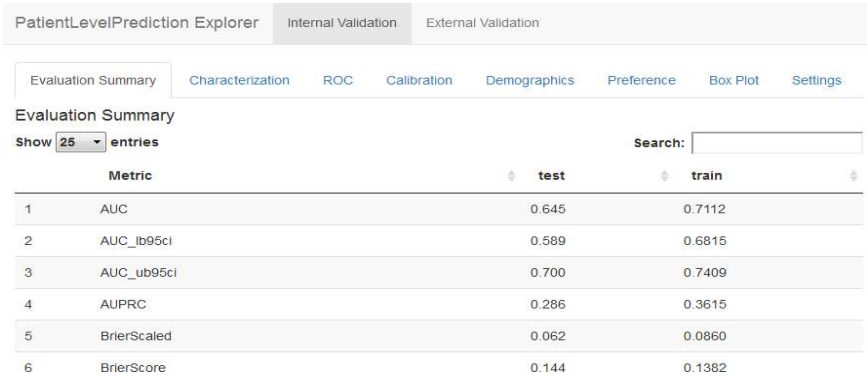
请注意，此示例为 LASSO 交叉验证和训练测试拆分设置随机种子，以确保多次运行的结果相同。

我们现在可以使用 Shiny 应用程序查看结果：

```
viewPlp(lassoResults)
```

这将启动

应用程序，如图 E.18 所示。在这里，我们在测试集上看到了一个 0.645 的 AUC，它比随机猜测要好，但对于临床实践而言有可能还不够好。



	Metric	test	train
1	AUC	0.645	0.7112
2	AUC_lb95ci	0.589	0.6815
3	AUC_ub95ci	0.700	0.7409
4	AUPRC	0.286	0.3615
5	BrierScaled	0.062	0.0860
6	BrierScore	0.144	0.1382

图 E.18：患者水平预测的 Shiny 应用程序

E10. 数据质量

练习 15.1

运行 ACHILLES:

```
library(ACHILLES)

result <- achilles(connectionDetails,

cdmDatabaseSchema = "main", resultsDatabaseSchema = "main",
sourceName = "Eunomia", cdmVersion = "5.3.0")
```

练习 15.2

运行数据质量看板:

```
DataQualityDashboard::executeDqCh
  ecks( connectionDetails,
  cdmDatabaseSchema = "main",
  resultsDatabaseSchema = "main",
  cdmSourceName = "Eunomia",
  outputFolder = "C:/dataQualityExample")
```

练习 15.3

查看数据质量检查列

```
DataQualityDashboard::viewDqDashboard( "C:/dataQualityExample/Eunomia/re
  sults_Eunomia.json")
```