ReDWINE：具备文本分析能力、助力康复研究的临床数据集市

David Oniani[1], Bambang Parmanto[1], Andi Saptono[1], Allyn Bove[3], Janet Freburger[3], Shyam Visweswaran[4,5,6], Nickie Cappella[4,6], Brian McLay[4,6], Jonathan C. Silverstein[4,6], Michael J. Becich[4,6], Anthony Delitto[3], Elizabeth Skidmore[2], Yanshan Wang, PhD[1,4,5,6]l卫生信息管理系，匹兹堡大学，美国宾夕法尼亚州匹兹堡；[2]职业治疗系，匹兹堡大学，美国宾夕法尼亚州匹兹堡；[3]物理治疗系，匹兹堡大学，美国宾夕法尼亚州匹兹堡； [4]生物医学信息学系，匹兹堡大学医学院，美国宾夕法尼亚州匹兹堡；[5]智能系统项目，匹兹堡大学，美国宾夕法尼亚州匹兹堡； [6]临床与转化科学研究所，匹兹堡大学，美国宾夕法尼亚州匹兹堡

Abstract

康复研究旨在确定治疗干预的组成部分、这些组成部分促成康复的机制，以及最终如何制定最佳干预策略，以最大限度地提升患者的身体、心理和社会功能。用于研究和确立新干预措施的传统随机临床试验面临诸多挑战，例如成本高昂、耗时较长。利用现有临床数据观察干预效果的观察性研究，相较随机对照试验（*RCT*）已显示出多项优势。电子健康记录（*EHR*）已成为开展观察性研究日益重要的资源。为支持此类研究，我们开发了一个临床研究数据集市，名为 *ReDWINE*（用于康复研究的信息学基础设施康复数据集市），该数据集市将 *UPMC* 医疗卫生系统收集的康复相关电子健康记录数据转换为观察性健康数据科学与信息学（*OHDSI*）观察性医疗结局合作组织（*OMOP*）通用数据模型（*CDM*），以促进康复研究。存储在 *ReDWINE* 中的标准化电子健康记录数据将进一步减少研究人员汇集、协调、清理和分析来自多个来源的数据所需的时间和精力，从而得出更稳健、全面的研究结果。*ReDWINE* 还部署了数据可视化和数据分析工具，以促进队列定义和临床数据分析。其中包括开放健康自然语言处理（*OHNLP*）工具包，以及一条高通量自然语言处理流程，可在 *ReDWINE* 中大规模提供文本分析能力。通过使用这一

ReDWINE 中全面呈现的患者数据开展康复研究，将有助于为健康干预及其结局提供真实世界证据。

介绍

根据世界卫生组织（WHO）的定义，康复是一系列旨在优化功能并减少急性和慢性疾病及状况患者残疾的干预措施，这些患者会与其环境发生互动（例如卒中、慢性疼痛、手术、癌症）”[1]。康复领域的研究人员采用随机临床试验（RCT），即循证医学的金标准，来研究变化机制，并为康复干预措施的有效性提供证据。然而，在康复研究中使用随机对照试验（RCT）存在若干挑战，包括研究结果缺乏普适性、成本高昂以及耗时较长 [2]。

观察性研究用于在不进行实验分配的真实世界环境中观察干预措施的效果。与随机对照试验（RCT）相比，观察性研究具有若干优势，包括外部效度更高，开展起来更快、成本更低。随着主要医疗卫生系统中电子健康记录（EHR）的出现，电子健康记录已成为开展观察性研究日益重要的资源 [3]。获取大型医疗卫生系统的电子健康记录数据对康复研究尤为有用，因为许多适合接受康复治疗的患者会在多个场所接受治疗。例如，因卒中入院的患者可能先在急症护理医院开始康复治疗，随后转至住院康复机构或专业护理机构继续康复，之后再转至社区，在家中、门诊或两者兼有的场所接受康复治疗。获取这些不同场所的电子健康记录数据，有助于全面了解患者的康复历程，并更准确地评估康复治疗的价值和效果。

为促进基于电子健康记录的观察性研究，观察性健康数据科学与信息学（OHDSI）社区开发了开源标准化词汇表和软件工具，例如观察性医疗结局合作组织通用数据模型（OMOP通用数据模型）。OMOP通用数据模型建立了一种标准化方法，用于将电子健康记录数据以及许多其他类型的数据（例如理赔数据、事务数据）映射并整合为一种通用格式，其中关联患者、测量结果、观察结果、就诊次数、医疗操作等。此外，OHDSI 社区还提供开源数据分析工具，以促进观察性研究。ATLAS [4] 是一款基于 Web 的工具，提供用户界面，用于可视化数据和标准化词汇表，并基于OMOP通用数据模型格式的观察数据定义队列。HADES [4] 是一款数据分析工具，提供一系列R软件包，用于大规模分析，包括人群特征描述、人群层面因果效应估计和患者层面预测。

电子健康记录中的非结构化叙述文本（如临床记录和操作记录）对康复研究至关重要。记录包含大量有关患者病史、症状、治疗和生活方式因素的信息，这些信息仅靠结构化电子健康记录无法获取 [5]。有些估计指出，临床记录中包含了研究所需的有价值数据，其中占比可达 70% 至 80%，甚至更高 [6]。例如，在一次物理治疗疗程中，只有临床记录才会记载特定运动的信息。对于有意研究不同运动与治疗结局之间关联的临床研究人员而言，这类信息很有价值。通过自动化提取非结构化电子健康记录信息的流程，临床与转化研究人员可以节省时间、提高效率，同时更全面地了解患者数据。OMOP通用数据模型通过新增 NOTE 和 NOTE_NLP 表，纳入了对非结构化电子健康记录的表示；这两个表分别存储临床文本数据和临床自然语言处理工具的输出。然而，OHDSI 社区目前尚未提供可从非结构化电子健康记录中提取医学概念的临床自然语言处理工具。

本文介绍一种名为 ReDWINE（用于康复研究的信息学基础设施康复数据集市）的临床研究数据集市原型的开发。该原型将大型医疗卫生系统收集的康复相关电子健康记录数据转换为 OMOP通用数据模型，以促进康复研究。存储在 ReDWINE 中的标准化电子健康记录数据能够减少数据协调与映射所需的时间和精力，并使研究人员能够汇集和分析来自多个来源的数据，从而得出更可靠、更全面的研究结果。我们还在 ReDWINE 中部署了 ATLAS 和 HADES 工具，以便进行队列定义和临床数据分析。此外，我们采用并部署了开放健康自然语言处理（OHNLP）工具包 [7]，——一个高通量自然语言处理流程，以大规模提供文本分析能力。利用 ReDWINE 中这种全面的患者数据表示开展康复研究，有助于为健康干预及其结局提供真实世界证据。

方法

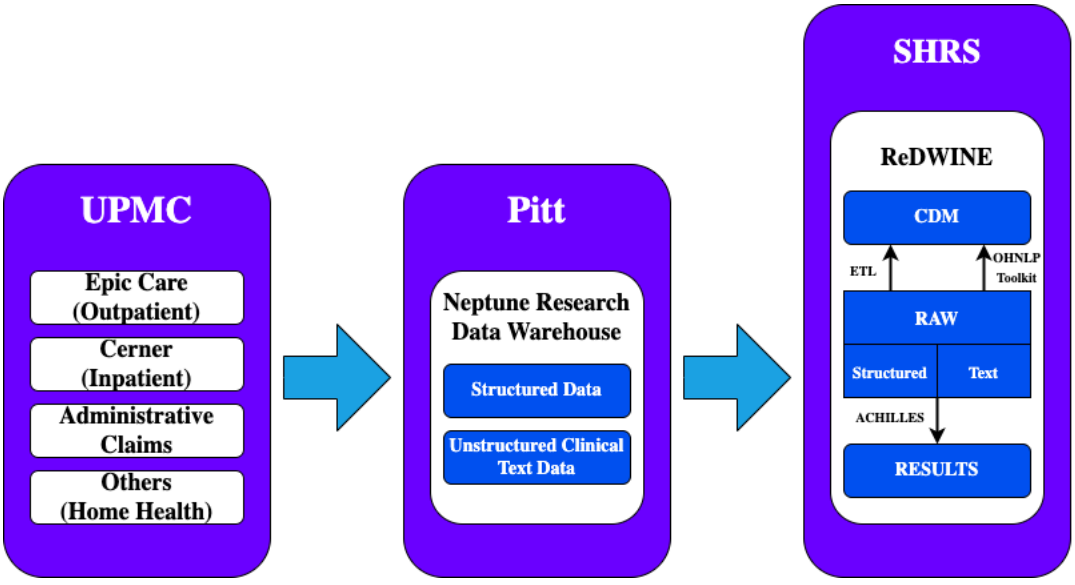

图*1. ReDWINE*架构与数据流程。

本节将介绍 ReDWINE 的实现细节。ReDWINE 是一个自助式信息学和数据分析平台，利用该大学的电子健康记录数据

匹兹堡医疗中心（UPMC），支持匹兹堡大学健康与康复科学学院（SHRS）的观察性康复研究。

数据收集

匹兹堡大学医学院生物医学信息学系设计并实施了一个名为 Neptune 的研究数据仓库 [8]，该仓库根据其与 UPMC 签订的业务伙伴协议（BAA），从 UPMC 提取、转换并存储电子健康记录数据。Neptune [8] 设有一个数据层，其中包含经过去标识化、符合 HIPAA有限数据集标准且保留日期的结构化数据，以及用于研究目的的非结构化电子健康记录临床文本数据。由于 UPMC 多种电子健康记录系统的复杂性，电子健康记录数据以最少的筛选和处理存储在 Neptune 中。ReDWINE团队与 Neptune 团队合作，从数据层提取与康复相关的电子健康记录数据，并将其转换为 OMOP通用数据模型，供康复研究人员使用。

在开发 ReDWINE原型 的第一步中，我们使用了经诊断为卒中的患者队列的电子健康记录数据。我们之所以选择卒中，是因为这是一种常见疾病，而且 SHRS 的许多康复研究人员都专注于卒中后的康复治疗。我们界定了一个患者队列，他们被诊断为卒中，并于2016年1月1日至2016年12月31日期间在 UPMC 急症护理医院收治（ICD-10代码见补充文件）。我们收集了这些患者卒中后12个月内在门诊环境中的所有可用电子健康记录数据，包括人口统计信息、诊断计费代码、操作代码、就诊记录和操作记录；这些记录创建于2016年1月1日至2018年12月31日之间。匹兹堡大学机构审查委员会（IRB）审查并批准了 ReDWINE 的开发（IRB编号21040204）。

数据库

在我们的实现中，我们使用了PostgreSQL[1] 数据库来构建 ReDWINE，因为它得到了OHDSI 社区的良好支持。PostgreSQL是一种免费且开源的对象-关系数据库管理系统（ORDBMS），专注于可扩展性和SQL 合规性。在该数据库中，我们定义了三个架构，即原始数据、通用数据模型和RESULTS。每个架构都有其自身的用途，并包含相应的表：

- RAW 架构：顾名思义，RAW 架构存储从 Neptune 获取的原始数据，这些数据随后会通过提取、转换、加载（ETL）流程进行转换、标准化和映射[2]。此架构基于从 Neptune 接收的电子健康记录数据，因此设计上可以灵活调整。该架构中的表取决于数据的性质和模态。

- CDM 架构：CDM 架构包含符合 OMOP 格式的标准化电子健康记录数据，并遵循 OMOP通用数据模型中定义的标准化词汇表和表。

- RESULTS 模式：RESULTS 模式包含由大规模纵向证据系统健康信息自动表征（ACHILLES）工具[3]生成的信息和统计数据。该工具用于验证数据质量，并确定数据是否符合指定要求。

由于稿件篇幅有限，ReDWINE 中各架构、ETL 流程以及文本分析工具的实现细节详见补充文件。

数据质量保证

由于数据在不同标准和代码之间进行了转换，我们开展了数据质量保证（QA）分析，以评估ETL 脚本的质量，并确保

[1] https://www.postgresql.org/ [2] https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html[3] https://github.com/OHDSI/Achilles

数据处理流程中没有发生重大数据丢失。具体做法是，分别统计 RAW 架构和 CDM 架构中原始数据的记录数，然后生成统计信息以评估数据丢失情况。此步骤可确保数据处理流程稳健可靠，并能妥善处理未来的数据流。

结果

*ReDWINE*原型统计

表1列出了使用卒中队列的 ReDWINE 原型统计数据。目前，数据库中共有 13,604 名患者，其中男性 6,673 名，女性 6,931 名。该队列共有 ~2 百万条诊断记录，平均每人 140 条诊断记录。经过自然语言处理后，ReDWINE 中该队列包含 ~2 百万条临床记录，共提取出 ~304 百万个医学概念，平均每位患者有 22,000 个医学概念。

表1. ReDWINE原型的队列统计。

	数量
患者	13,604
Male	6,673
女性	6,931
诊断	1,909,175
人均诊断次数（平均）	140
就诊次数	1,987,791
人均就诊次数（平均）	146
医疗操作	1,214,388

每位人员的医疗操作（平均值）	89
备注	2,096,415
每名人员的备注（平均值）	154
临床自然语言处理概念（注释_自然语言处理表)	304,228,932（60,627 唯一）
每人概念数（平均）	5,018

数据质量保证结果

表2展示了对 RAW 和 CDM 架构记录数进行数据质量分析的结果。我们发现，数据转换 ETL 脚本在处理人口统计信息、诊断、就诊次数和医疗操作时表现良好，数据丢失极少（范围为 0% 至 9.627%）。对于 RAW.demographics 与 CDM.person 的比较，没有数据丢失，RAW 架构中的表内每位人员都映射到了 CDM 架构中的相应表。对于 RAW.diagnoses 与 CDM.condition_occurrence 的比较，有 2,475 条记录未能映射（0.129%）。对于 RAW.encounters 与 CDM.visit_occurence 的比较，所有记录均成功映射到 CDM 表中，没有数据丢失。最后，对于 RAW.procedures 与 CDM.procecure_occurrence 的比较，由于必须排除若干不在 OMOP 通用数据模型标准化词汇表中的 CPT- 4 代码，我们未能映射 129,371 条记录。因此，原始 1,343,759 条记录中有 9.627% 被排除。总体而言，RAW 和 CDM 架构之间的记录数一致，表明数据已成功传输至 OMOP 通用数据模型。

表2。ReDWINE原型的数据质量保证分析结果。

比较		RAW	CDM	差异 e	原始数据丢失 (%)
原始数据.人口统计信息 vs 通用数据模型.人员		13,604 1		0	0
原始数据.诊断 通用数据模型.状况 _发生	vs	1,911,650	1,909,175	2,475	0.129
RAW.encounters CDM.visit _发生	vs	1,987,791	1,987,791	0	0
RAW.procedures CDM.操作_发生	vs	1,343,759	1,214,388	129,371	9.627

案例研究 *1*：使用 *ATLAS* 定义队列并可视化结果

假设某位康复研究人员有兴趣确定一个患者队列，这些患者接受过物理治疗，并且先被诊断为 2 型糖尿病，随后又被诊断为卒中，那么他/她可以使用 ATLAS 中的队列定义工具设定这些标准。研究人员可以使用 OHDSI Athena 工具查找标准化的 OMOP CDM 概念 ID，以便为纳入标准定义三个概念集。利用定义好的概念集，研究人员可以在 ATLAS 中定义队列，如图 2 所示。队列进入事件被定义为队列索引时间，即数据库中的受试者进入队列的时间。在本例中，我们使用卒中诊断（概念 ID：35207821）作为进入事件。随后，研究人员可以定义纳入标准，进一步筛选队列。在本案例研究中，我们使用了物理治疗（概念 ID：2314284）和 2型糖尿病（概念 ID：35206882）。也可以在每个概念集中添加更多相关概念 ID，不过为了演示，我们每个概念集只使用了一个概念 ID。接下来，研究人员可以.....

研究人员可以在“生成”功能中生成队列。在本案例研究中，我们能够识别出五名符合队列标准的患者。

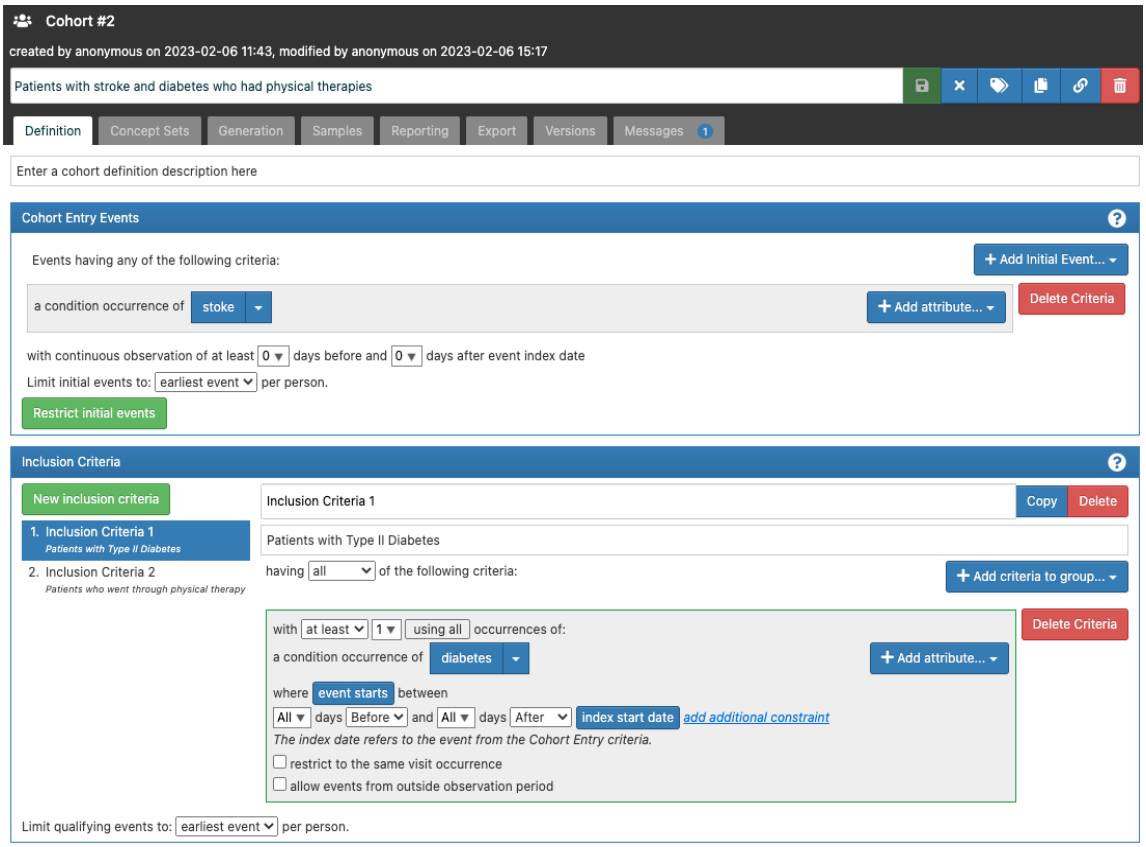

图 2。 Atlas 中的队列定义。

案例研究 *2*：利用自然语言处理提取的概念识别患者

非结构化电子健康记录中的临床信息对临床与转化研究人员可能非常重要。ReDWINE 中通过自然语言处理提取的概念可促进此类信息在观察性研究中的应用。在一项案例研究中，我们假设一名研究人员希望识别经历过一次或多次跌倒的患者。我们使用 Note_NLP 表和 note_nlp_concept_id “436583”来识别临床记录中提及跌倒事件的患者。我们使用 HADES 工具识别出 10,322 名患者。我们将这些患者的数据载入

R 数据框以便进一步分析。我们还发现，Note_NLP 表中记录在临床记录里的跌倒事件共有 358 种不同的词汇变体。例如，“accident &fell”“Accident a fall”“accident and fall”“accident and falling”“accident and fell”“accident from falling”“accident/ fall”“accidents/falls”“down Falls”“down Fall”“down and falling”“down and falls”“fall”和“fall and injured”。用于识别这些患者的 R 脚本可在补充文件中找到。在另一项研究中，[9], 我们利用 ReDWINE 中的 OHNLP 工具包，开发了一种定制的基于规则的自然语言处理算法，用于从临床记录中提取物理康复运动信息。物理康复运动信息涵盖九个类别中的 101 个医学概念，包括动作类型、身体侧别、身体部位、运动平面、持续时间、组数和重复次数信息、锻炼目的、锻炼类型以及身体姿势。ReDWINE 提供的电子健康记录文本分析能力，是朝着实现精准康复承诺迈出的重要一步[10]。

讨论

随着电子健康记录数据量不断增加、计算能力不断提升，获取并利用近乎实时的数据开展研究和临床工作已成为现实。我们介绍了一种方法，可对来自多个场景、采用不同电子健康记录软件（如 Epic、Cerner）的电子健康记录数据进行协调与标准化，并以卒中康复护理为应用示例。精准康复近来在研究和实践中都受到关注，其目标是在恰当的时间提供精准且个体化的干预 [11]。ReDWINE 开发的信息技术基础设施、标准化数据元素、通用数据模型和数据分析平台，以及借助 ReDWINE 开展的观察性研究，都可用于推动精准康复研究，最终帮助急性疾病或损伤患者以及慢性状况患者最大限度地恢复功能、减少残疾。此外，ReDWINE 还提供了一个平台，用于

开展学习型卫生系统研究，旨在将卫生系统数据和经验与外部证据相结合，以改进实践。因此，ReDWINE 支持建立这样的体系，使患者能够获得质量更高、更安全、更高效的护理。

能够随时间推移并跨不同医疗环境追踪患者，对于康复研究尤其重要，因为康复护理通常跨不同医疗环境开展，并持续数月。然而，当需要了解患者随时间推移的结局和医疗服务利用情况时，这一功能还具有更广泛的临床和卫生系统意义。例如，在捆绑支付和责任制医疗组织等支付模式下，卫生系统会获得一笔总额支付，用于在特定时期内管理患者。了解患者在该时间段内的医疗服务利用轨迹，可以提供有用信息，从而改善医疗服务的质量和价值。

在 ReDWINE 中纳入自然语言处理工具，使康复研究人员能够在比没有自然语言处理工具时大得多的程度上研究特定干预措施的效果。康复服务提供者所实施的具体运动及其他干预措施，通常只能在患者记录中的自由文本临床记录里找到。因此，仅查看一次就诊期间计费的 CPT代码，通常无法提供研究人员确定某项特定干预措施效果所需的详细程度。例如，物理治疗师可能会在一次治疗期间为三项每项15分钟的“治疗性运动”计费。仅使用结构化数据字段，通常无法得知进行了哪些运动。然而，借助自然语言处理工具，研究人员可以提取有关所实施的具体运动及剂量（组数、重复次数、频率、持续时间）的数据，如 [9] 所示。这将使研究人员能够更加精确地确定寻求康复护理的个人的最佳治疗策略。

ReDWINE利用OMOP通用数据模型和标准化词汇表这一特点，对于康复电子健康记录数据至关重要，因为UPMC旗下的多家医院使用不同供应商提供的多个电子健康记录系统[8]。例如，不同的UPMC居家健康服务使用不同的

电子健康记录系统。尽管 UPMC 正在转向集成式 Epic 电子健康记录系统，但整合来自多个系统的历史电子健康记录数据，对于康复研究仍至关重要。ReDWINE 能够整合并协调来自多个来源的电子健康记录数据，这些数据采用不同的数据格式和编码系统；它还能解决不同来源之间的差异，以确保数据的一致性和准确性。ReDWINE 提供了一个灵活且可扩展的框架，使康复研究人员能够开展观察性研究、卫生服务研究和质量改进工作，并促进多中心研究。

尽管 ReDWINE 主要面向观察性研究和卫生服务研究开发，但其存储全面电子健康记录数据的能力也可用于评估 RCT 的可行性 [12]，并有助于在本地卫生系统中招募代表性不足的人群。众所周知，临床试验中来自种族和族裔少数群体的代表性不足人群招募率较低，这仍然是一个问题，并已造成许多健康差异。利用 ReDWINE 汇集的来自多个来源的丰富人口统计信息，有望帮助识别适合参与 RCT 的代表性不足人群。

局限性

ReDWINE 存在若干局限性。首先，ReDWINE 仍是一个原型，使用来自单一数据源的小规模电子健康记录数据集实现。因此，该系统尚未在生产环境中进行分析，无法体现拥有大量用户和持续数据流的情境。尽管如此，我们已在 ReDWINE 中实现了所有必要的组件和功能，使其能够存储更多数据。其次，由于我们希望使该研究易于访问和复现，对于所有主要的数据分析和探索工具，我们都依赖 OHDSI 框架及其工具（例如 ATLAS、HADES）。因此，对于定制的、非 OHDSI 工具的开发和使用尚未进行充分探索。最后，尽管 OMOP通用数据模型 对观察性研究具有若干优势，但它并不是唯一的标准化通用数据模型。因此，ReDWINE 也可能存在 OMOP通用数据模型 中已有的局限性。

未来工作

我们已经使用卒中患者队列的电子健康记录数据构建了 ReDWINE 原型。该原型具备促进康复研究所需的功能。目前，该原型仍处于早期阶段，旨在为匹兹堡大学 SHRS 的多位研究人员提供服务。当前的一项紧要任务是邀请几位测试用户使用 ReDWINE 开展观察性研究，测试队列定义和数据分析功能，并征求他们的反馈以改进原型。要将 ReDWINE 打造成供研究人员使用的自助工具，还有几项工作需要考虑。首先，我们将与机构 IRB 合作，确保研究人员在使用 ReDWINE 前获得适当的 IRB 批准。我们将实施审计和隐私设置，确保研究人员能够充分发挥系统的潜力。其次，我们将制定服务与维护商业模式，确保 ReDWINE 支持团队能够持续运作。第三，我们将与 Neptune 团队合作，确保电子健康记录数据能够定期更新。第四，我们将为 HADES 工具中常用的数据检索功能开发若干 R 封装函数。这将使没有 SQL 使用经验的研究人员更方便地使用数据分析工具。最后，随着人工智能（AI）模型和技术的快速涌现，将此类能力纳入 ReDWINE 的数据分析工具颇具意义。尽管目前已能够自动执行统计分析和 AI 模型训练（例如自动化机器学习 [13]），但这些功能尚未集成到 ReDWINE 中。因此，构建 AI 模型训练的数据处理流程或许是未来工作中值得探索的另一个方向。

作者贡献

DO：设计；撰写手稿，BP：构思研究；编辑手稿，AS：编辑手稿，AB：编辑手稿，JF：编辑手稿，SV：构思研究；编辑手稿，NC：获取数据；编辑手稿，BM：获取数据，JCS：构思研究；编辑手稿，MJB：构思研究；编辑手稿，AD：构思研究；编辑手稿，ES：构思

该研究；编辑了稿件，YW：构思了该研究；分析了数据；撰写了稿件。所有作者均阅读并批准了最终稿件。

致谢

作者谨此感谢 OHDSI 社区提供详尽的通用数据模型（CDM）文档，并发布 OMOP 数据分析工具。本项目由匹兹堡大学健康与康复科学学院和美国国立卫生研究院通过资助编号 UL1TR001857 和 U24TR004111 提供支持（SV、NC、BM、JCS、MJB、YW）。Neptune 数据仓库由美国国家转化科学促进中心、美国疾病控制中心、美国国家职业安全卫生研究所、美国国立卫生研究院共同基金以及患者中心结局研究所临床研究网络提供支持（UL1TR001857、RI-CRN-2020-006、OT2OD026554 和 5U24 OH009077——资助对象为 MJB、JCS、SV）。资助方未参与研究设计、数据收集、分析和解释，也未参与论文撰写。本报告所述观点不一定代表资助方的观点，仅代表作者个人观点。

利益冲突声明

作者声明不存在利益冲突。

总结表

- 我们开发了一个名为 ReDWINE（用于康复研究的信息学基础设施康复数据集市）的临床研究数据集市，将从本地医疗卫生系统收集的康复相关电子健康记录数据转换为观察性健康数据科学与信息学（OHDSI）观察性医疗结局合作组织（OMOP）通用数据模型（CDM），以促进康复研究。

- ReDWINE 集成了数据可视化和数据分析工具，以便进行队列定义和临床数据分析。

- ReDWINE 通过实施高吞吐量的自然语言处理流水线——开放健康自然语言处理（OHNLP）工具包，提供大规模文本分析能力。

- 数据质量保证分析表明，电子健康记录数据已成功转移至 ReDWINE 中的 OMOP通用数据模型。

- 两个案例研究展示了 ReDWINE 在康复研究中的应用。

参考文献

1 [_____] .
2

https://www.who.int/news-room/fact-sheets/detail/rehabilitation 。Kroll T, Morris J：在康复研究中使用混合方法设计的挑战与机遇。《物理医学与康复档案》2009，90：S11-S16。3. Callahan A, Shah NH, Chen JH：使用电子健康记录数据开展观察性研究的研究与报告注意事项。《内科医学年鉴》2020，172：S79-S84。4. OHDSI：《OHDSI 之书：观察性健康数据科学与信息学》。OHDSI；2019。5. Murphy SN, Visweswaran S, Becich MJ, Campion TR, Knosp BM, Melton-Meaux GB, Lenert LA：研究数据仓库最佳实践：通过信息学创新推动国家级数据共享。第29卷。第581-584页：牛津大学出版社；2022：581-584。6. Becich M：信息管理：从检测结果迈向临床信息。《临床领导力与管理评论：CLMA 期刊》2000，14：296-300。7. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, Liu S, Sohn S, Liu H, Fan J：加速医疗人工智能发展的自然语言处理服务设计原则及梅奥诊所的自然语言处理即服务实施。《NPJ 数字医学》2019，2：130。8. Visweswaran S, McLay B, Cappella N, Morris M, Milnes JT, Reis SE, Silverstein JC, Becich MJ：研究数据仓库设计与实施的原子化方法。《美国医学信息学会杂志》2022，29：601-608。9. Shaffran SW, Gao F, Denny PE, Aldhahwani BM, Bove A, Visweswaran S, Wang Y：从临床记录中提取物理康复运动信息：基于规则与机器学习自然语言处理技术的比较。arXiv 预印本 arXiv:230313466，2023。10. Cotton RJ, Segal RL, Seamon BA, Sahu A, McLeod MM, Davis RD, Ramey SL, French MA, Roemmich RT, Daley K：精准康复：优化功能，为医疗保健创造附加价值。《物理医学与康复档案》2022，103：1883-1884。11. French MA, Roemmich RT, Daley K, Beier M, Penttinen S, Raghavan P, Searson P, Wegener S, Celnik P：精准康复：优化功能，为医疗保健创造附加价值。《物理医学与康复档案》2022，103：1233-1239。12. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M：OMOP通用数据模型可助力癌症预测的数据驱动研究：系统综述。《国际分子科学杂志》2022，23 11834：。13. He X, Zhao K, Chu X：自动化机器学习：最新进展综述。《知识型系统》2021，212 106622：。

补充文件

A. ReDWINE 中各架构的实现细节以及提取、转换和加载（ETL）流程。

原始数据架构：

在 ReDWINE 的实现中，原始数据架构包含以下表：

- 人口统计信息：包含患者标识符、出生日期、死亡日期、性别、种族和族裔等信息。

- 诊断：包含患者标识符、诊断类型、诊断代码（即 ICD 代码）、诊断名称和诊断日期等信息；

- 就诊：包含患者就诊数据，例如就诊标识符、就诊类型和就诊日期；

- 就诊记录：包含通过NLMscrubber 进行清理、以尽力实现去标识化的就诊记录（仍包含一些 PHI）以及相关就诊元数据（患者标识符、就诊标识符、就诊日期等）；

- 医疗操作：包含与医疗操作相关的数据，例如操作标识符、操作类型、操作代码；

- 操作记录：包含经 NLMscrubber 清理、尽力去标识化的操作记录（仍包含部分受保护健康信息），以及相关操作元数据（患者标识符、操作标识符、操作日期）。

我们在 RAW 架构中创建了这些表，然后使用 SQL 编程语言[1]将从 R3 收到的电子健康记录数据加载到这些表中。对于结构化电子健康记录数据，我们使用了未经处理的原始数据。对于非结构化电子健康记录数据（就诊记录和操作记录），一

[1] https://www.postgresql.org/docs/current/sql.html

次就诊或一次操作可能在原始格式中包含多个文本数据条目。因此，我们运行了 SQL 预处理脚本，将每次就诊和每项操作的自由文本条目合并起来。事实证明，这对我们用于从记录中自动提取概念并将其映射到标准化概念标识符的临床自然语言处理工具非常有帮助。我们将在下一节详细介绍该工具。

通用数据模型架构：

CDM 架构包含来自多个来源的数据，包括从原始数据架构获取的电子健康记录数据、OMOP 的标准化词汇表以及相关元数据。该架构使用 OHDSI 社区 OHDSI/CommonDataModel[2] 提供的 CDM R 包生成。如图 A1 所示，此过程包括三个步骤：1）使用该工具生成数据定义语言（DDL）SQL 程序；2）以自动化方式对 DDL 程序进行进一步修改；3）在 ReDWINE 中运行这些程序。

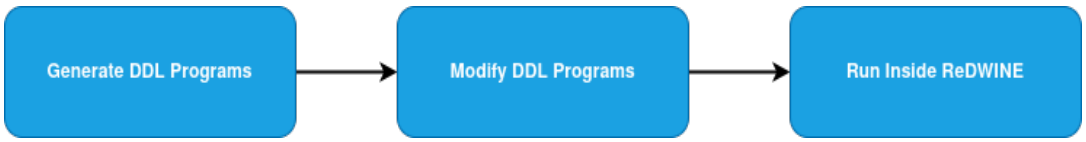

图 *A1. DDL* 生成与运行流程

此 CDM R 包可用于生成各种版本的OMOP通用数据模型 DDL。我们使用的是CDM5.3版。原因在于所有主要的 OHDSI 工具都支持此CDM 版本。根据数据的不同，DDL 脚本可能需要进一步修改。就我们的情况而言，我们必须增加文本文件允许的最大长度，并更改整数类型。之所以需要更改文本文件大小，是因为我们的一些文件大于默认最大值。至于整数类型的更改，我们必须修改 DDL 程序，因为默认的整数

[2] https://github.com/OHDSI/CommonDataModel

类型无法容纳用于表示人员标识符和注释标识符的大数值。我们通过预先定义所有这些替换项并将其整合到一个 R 程序中，实现了这些 DDL 程序转换的自动化。所有步骤成功完成后，便会在 ReDWINE 中运行 DDL 程序，以创建 CDM 架构，以及 OMOP通用数据模型所定义的所有 CDM 架构表。

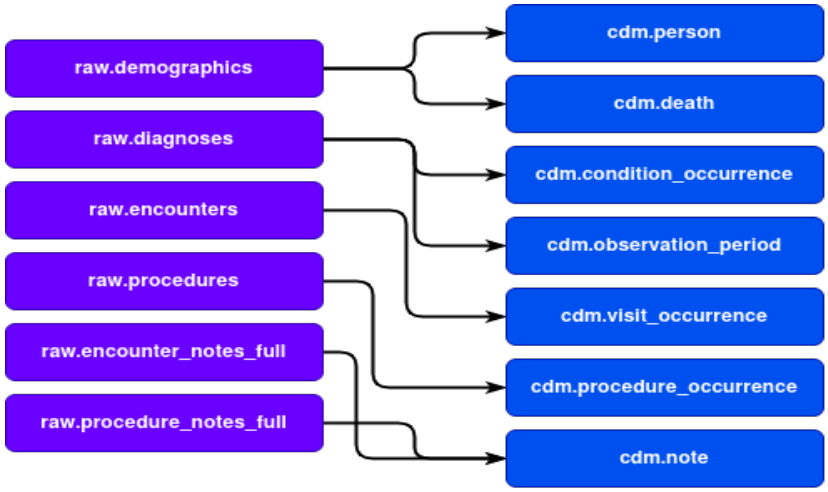

图 *A2*。表：从 *RAW* 架构到 *CDM* 架构的映射

ETL 流程是数据集市中最重要的流程，应当谨慎执行。图 A3 展示了 ETL 流程的流程图。CDM 架构包含在 ETL 流程中填充数据的表。在 ETL 流程中，我们从 RAW 架构提取数据，对其进行转换，然后加载到 CDM 架构中。对于我们的试点数据集，我们采用了从 RAW 架构到 CDM 架构的直接映射，如图 A2 所示。这种映射不一定是一对一的，也就是说，某些表（例如人口统计信息）可以用于填充两个不同的 CDM 表。映射或数据加载流程几乎总是需要一些手动操作，以便将特定于数据的概念语言转换为符合 OMOP CDM 标准的语言。不过，这项工作一旦完成，其余步骤便可自动化。

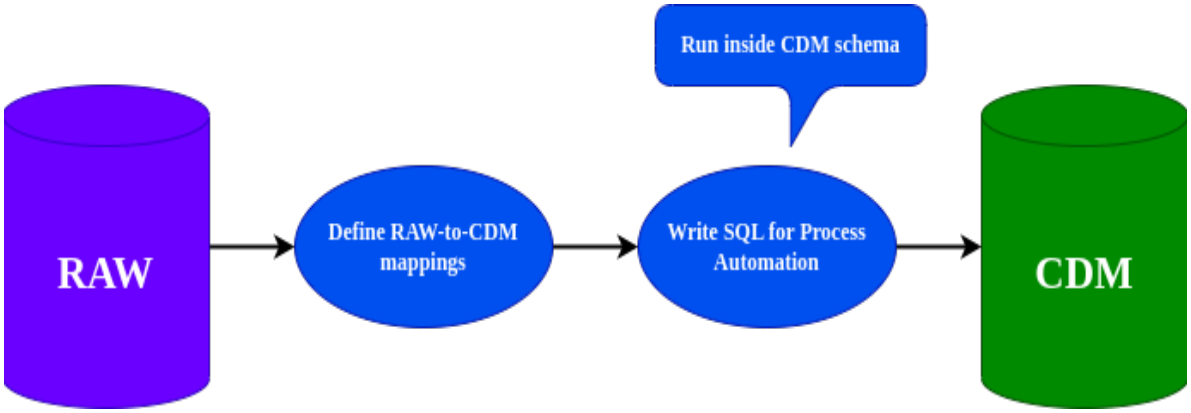

图 *A3*。*ETL* 流程。

加载流程首先要将 RAW 架构中的数据映射到 OMOP CDM 支持的标准化词汇表。此步骤通常采用自动映射和手动映射两种方法。对于遵循医疗领域广泛使用标准（如 ICD-9/10、CPT-4 等）的代码，可以采用自动映射。在自动映射步骤中，我们使用了从 OHDSI Athena[3] 中请并获取的

加载到 CDM 架构中。然后应用自行开发的 SQL 函数执行自动代码到代码映射。该函数利用加载到数据库中的字典，将源数据中的代码映射到目标 CDM 概念标识符。在我们的案例中，诊断的

下：to_cdm(cpt_code, 'CPT4')。其中源代码来自 CPT 词汇表，目标则为 OMOP CDM 概念 ID。我们开发的 SQL 函数还支持多词汇表查找，在这种情况下，它会尝试确定源概念 ID 是否存在于

[3] https://athena.ohdsi.org/

10 个词汇。在这种情况下，该函数会首先检查代码是否来自 CPT-4 词汇；如果查找失败，则会转而查询 ICD-10 词汇，并返回相应的 OMOP CDM 概念 ID。

当源概念无法映射到 OMOP CDM 标准化词汇表时，我们会在康复领域专家的协助下进行手动映射。如果医疗卫生系统采用了自己的命名约定，就可能出现这种情况。实际操作中，可在 SQL 中使用 SQL 的 CASE 语句[4]或类似方式，针对个别情况进行判断，并将每种情况映射到

HISPANIC OR LATINO”和“NOT HISPANIC OR LATINO”）与相应的 OMOP CDM 概念标识符（分别为 38003563 和 38003564）之间建立映射。表 A1 列出了负责处理此映射的 SQL代码示

卫生系统和临床环境，有时还包含地点名称。领域专家协助查找 OMOP CDM 标准化词汇表中语义上最接近的概念。如果找不到相应概念，我们就按照 OHDSI 的建议，将概念标识符设为 0。

SQL代码示例。

表 A1。处理族裔映射的 SQL代码示例。

```
1 CASE 2 WHEN ethnicity = '西班牙裔或拉丁裔' THEN
38003563 3 WHEN ethnicity = '非西班牙裔或拉丁裔' THEN
38003564 4 ELSE 0 5 END AS ethnicity_concept_id
```

[4] https://www.postgresql.org/docs/current/functions-conditional.html

加载标准化数据后，我们使用 OHDSI 的 ACHILLES 工具进行了一系列分析，以确保数据质量。ACHILLES 可在 CDM 架构上运行，为 OMOP CDM 数据库生成描述性统计数据。这些统计数据随后可供 ATLAS[5] 等数据分析工具使用。统计数据本身会加载到 RESULTS 模式中。

表 A2。将 RAW 架构中的 enc_type 列映射到 CDM 架构中的 OMOP CDM 概念 ID 的 SQL代码示例。

```
1 CASE
2 WHEN enc_type = '摘要' THEN 45877039
3 当 enc_type = '过敏原混合物' THEN 44791052
4 WHEN enc_type = '过敏针剂' THEN 46272888
5 当 enc_type = '门诊就诊摘要' 时，则为 46237886
6 WHEN enc_type = 'ANTICOAGULATION' THEN 35803400
7 WHEN enc_type = '已安排抗凝治疗' THEN 35803400
8 WHEN enc_type = '预约' THEN 4089197
9 --省略了100多个 WHEN 语句
10 ELSE 0
11 END AS 就诊_概念_ID
```

文本分析能力

完成 ETL 流程并在数据上运行 ACHILLES 工具后，我们使用临床自然语言处理工具从就诊记录和操作记录中自动抽取医学概念，并将其映射到相应的 OMOP 通用数据模型标准化概念标识符。我们使用 OHNLP 联盟开发的 OHNLP 工具包，自动从临床记录中抽取概念

[5] https://github.com/OHDSI/Atlas

抽取。OHNLP 工具包采用基于本体派生词典的方法，并以 MedTagger [1] 作为信息抽取的基础。具体而言，系统从记录中抽取医学概念，将其规范化为统一医学语言系统（UMLS）的概念唯一标识符（概念唯一标识符），随后映射到相应的 OHDSI 标准化概念标识符，最后存储在 Note_NLP 表中。系统使用精确字符串匹配算法自动完成映射；如果找不到完全匹配的文本，则进行人工链接。使用 OHNLP 工具包的另一项优势是，用户可以定义自定义规则，从记录中抽取特定概念和表型。该流程会将数据填充到 CDM 架构中的 Note NLP 表。_

OHNLP 工具包使用 Apache Flink——一个开源、统一的流处理和批处理框架——来监测批量文本数据处理过程及相应的统计信息，如图 A4 所示。我们发现，这一功能对于调试极其有帮助。具体而言，对于 ReDWINE，初始的“任务堆”内存不足，导致程序无法成功退出。该工具帮助我们调试了这一问题，随后我们将内存增加到 52.0 GB。OHNLP 工具包耗时约 8 小时，完成了对超过 200 万条记录的文本分析和概念抽取，并将这些概念填充至 CDM 架构的 Note_NLP 表中

。关于 OHNLP 工具包的更多详情，请参见之前的研究[2]。

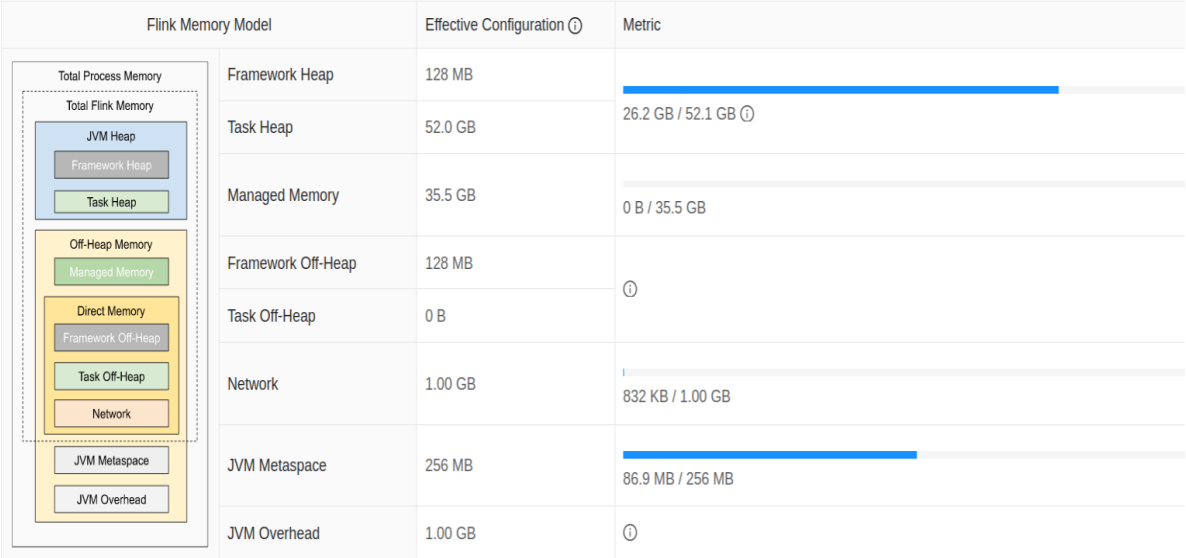

图 *A4.*通过 *OHNLP* 工具包在服务器上运行的 *Apache Flink*。

数据分析工具

Broadsea 是一款工具，可在 Docker 容器[6]中运行 OHDSI 核心数据分析技术栈。其中包括 ATLAS 网络应用程序，以及在 RStudio Server 中运行的 HADES R 软件包。核心技术栈的容器化支持跨平台部署。这意味着 ReDWINE 可以部署在大多数主流平台上，包括 Linux、macOS 和 Windows。在我们的案例中，我们使用 Linux 来托管数据库并部署 Docker 容器。值得注意的是，OHDSI 社区广泛支持 Broadsea 2.0，因此我们在其发布几周后便迁移到了该版本。总体而言，迁移过程十分顺利，但仍需配置数据库连接，以便 Broadsea 2.0 访问通用数据模型（CDM）数据以及由 ACHILLES 填充的结果表。部署

[6] https://www.docker.com/resources/what-container/

Docker 容器后，我们便在容器中运行 ATLAS 和 HADES，研究人员可以通过部署在 ReDWINE 上的 Broadsea 2.0 门户访问二者。

参考文献

1. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, Ravikumar K, Wu ST, Kullo IJ, Chute CG: **利用电子健康记录进行队列识别的信息抽取框架**。*AMIA* 转化科学峰会论文集 2013年，**2013年：**149. 2. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, Liu S, Sohn S, Liu H, Fan J: **加速医疗人工智能发展的自然语言处理交付愿景及梅奥诊所的自然语言处理即服务实施**。*NPJ* 数字医学 2019年，**2：**130。