

卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心

2024. 7. 24

（七） 数据仓库

1. **标题：**数据湖、数据仓库、数据集市和功能存储：对完整数据重用途径的贡献

来源：JMIR Med Inform

时间：2024 年 7 月.

链接：<https://www.jmir.org/10.2196/54590>.

概要：随着健康信息技术的日益普及和使用，产生了大量电子格式的临床数据，为数据重用提供了更多机会，甚至超过直接患者护理。然而，由于数据分布在多个软件中，格式、用词和技术的差异以及软件之间缺乏统一的标识，使得在不同数据源之间交叉引用信息成为大问题。为了应对这一难题，医院采用了数据仓库来整合和标准化这些数据以进行研究。此外，作为补充或替代，数据湖以详细且未经处理的格式存储源数据和元数据，从而能够对数据进行探索、操作和调整，以满足特定的分析需求。随后，数据集市用于进一步将数据细化为针对特定研究问题量身定制的可用信息。但是，为了进行高效分析，特征存储对于透视数据和非规范化数据、简化查询至关重要。总之，虽然数据仓库至关重要，但数据湖、数据市场和特征存储在促进医疗保健研究和分析的数据重用方面发挥着重要和互补的作用。本研究提出了数据重用途径组件的标准化命名法和定义。虽然数据仓库是必要的初始阶段，但数据市场和特征存储的集成增强了其有效性。数据市场从原始数据中计算相关信息，而特征存储则将其组织到列中，从而简化数据集构建。而数据湖成为将原始数据存储存储在单个位置的宝贵资源。但值得注意的是，在没有数据湖的数据途径中，在分析之前完成 ETL 过程的要求会带来延迟。研究中描述的特定组件及其特征不是一成不变的，可以根据独特的组织需求和配置而有所不同。例如，数据湖和特征存储的包含通常是自由裁量的，受源数据的规模和复杂性、特征的数量、研究项目的范围、团队的规模以及研究可重复性随时间推移的必要性等因素的影响。

2. **标题：**在法国地区大学医院中实施从蓝图到临床的生物医学数据仓库：揭示流程、克服挑战并提出临床见解

来源： JMIR Med Inform.

时间： 2024 年 6 月.

链接： <https://doi.org/10.2196/50194>.

概要： 生物医学数据仓库（BDW）已成为促进健康数据在研究和决策应用中重用的重要工具。除了技术问题之外，BDW 的实施还需要强大的机构数据治理以及欧洲和法国法律框架的运营知识，以管理研究数据的获取和使用。本研究主要描述了在法国地区大学医院里综合实施 BDW 的过程和内容。文章首先描述了为开发南特 BDW 而进行的有关组织变革、技术架构和共享治理的行动和挑战，然后是访问临床内容的过程，并提供了有关患者数据保护的详细信息，还使用了示例来说明合并临床见解。文章还从 BDW 实施的法律职能框架、信息技术管理和互操作性、科学数据管理和对 BDW 的代理访问等几个方面进行了详细论述，特别是数据管理和代理访问，又具体对其目标和数据访问、数据保护和患者同意、监管许可、预算规划和可持续性和伦理进行了分析。本研究选取的数据是 2023 年 5 月 10 日南特大学医院生物医学数据仓库数据，涉及 2003–2022 年约 150 万名 CHUN 患者的 6800 多万份文本文件和 5.43 亿份编码信息，可以查询和转换，以提供给调查人员。南特 BDW 自 2018 年成立以来，已有 269 个项目根据数据使用和法规要求对数据进行访问，其中 115 个项目创建了数据集市。最后研究得出的结论是临床数据仓库的发展为来自临床护理的大量不同数据提供了前所未有的访问，但是它还需要在治理、与患者同意和数据保护相关的数据访问规则以及技术挑战方面付出一定努力。在功能研究框架内重组结构和多学科团队合作可能是 NBDW 成功要素，进一步推进质控过程和建立网络共同努力可能是需要吸取的最好教训。未来的挑战是建立标准化和共享的质控办法，以确保区域共享中数据质量，扩大信息技术和数据方面的长期投资，加速 BDW 的使用，而且数据治理必须优先于机构数据战略中对护理专业人员和患者的技术问题。

（徐健编辑）

译文一：

数据湖、数据仓库、数据集市和功能存储：对完整数据重用途 径的贡献

Antoine Lamer, Chloé Saint-Dizier, Nicolas Paris, Emmanuel Chazard, 徐健(译)

介绍

在过去几十年中，卫生信息系统（HIS）的广泛采用和使用已将大量临床数据从手动格式转变为电子格式。HIS 为了医护、管理或计费而收集和提供数据。除了这些初始用途外，HIS 还提供数据重用的机会，即“个人健康信息的非直接医护使用”，例如研究、医护质量、活动管理或公共卫生。医院已逐渐采用数据仓库来促进数据的重用。即使数据仓库是一个新流行的概念，数据重用也不限于提供和查询数据仓库。从这个角度来看，我们的目标是概述数据重用途径的不同组成部分，以及它们如何相互补充和互连。这个定义来源于我们通过与不同机构的同事合作获得的个人经验和见解。此外，我们借鉴了该领域专业人士分享的集体经验，有助于全面理解不同医疗保健环境中的数据重用实践。途径如图 1 所示并有详细说明。表 1 比较了每个组件的特性。最后多媒体附录 1 提供了数据重用途径的每个组件的数据、结构和体系结构的示例。

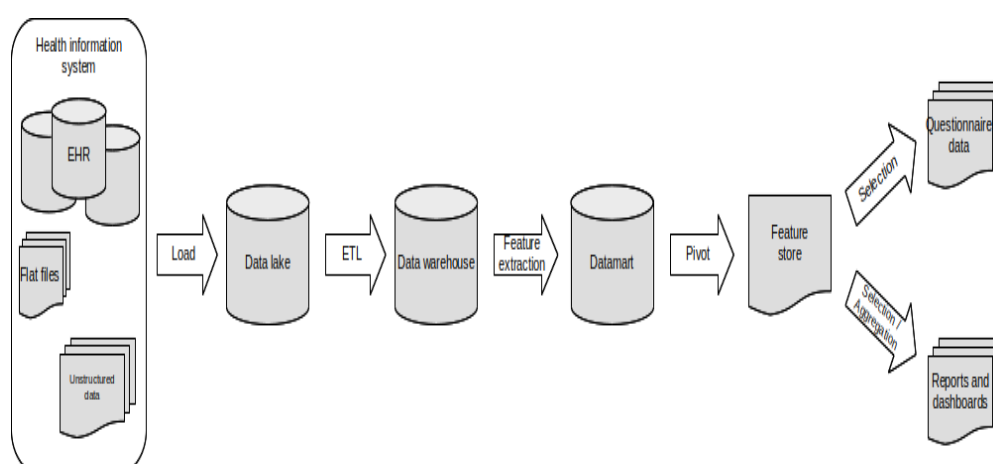


图 1. 用于数据重用完整途径的组件（EHR：电子健康记录；ETL：提取-转换-加载）

表 1 数据重用途径的每个组件的特征

特性	软件	数据湖	数据仓库	数据市场	特征存储
内容	数据和元数据	数据和元数据	数据	特征	功能和功能元数据
架构	分散式	集中	集中	集中	集中
内容水准	细粒度	细粒度	细粒度	聚合	聚合
数据	原始数据	原始数据	清洗数据	清洗数据	清洗数据
命名	多样化	多样化	标准化	标准化	标准化
数据模型	规范化	规范化	规范化	规范化	非规范化
数据结构	面向行	面向行	面向行	面向行	面向列
目的	事务性软件	特别探索性查询	所有用途	预先指定用途	预先指定用途

伦理考量

本研究不包括人类参与者研究（未进行人类参与者、实验或干预），因此不需要机构审查委员会的批准。

健康信息系统

存储在 HIS 中原始数据分布在多个软件中，由于数据格式的变化，从表格到层次结构再到自由文本，因此无法在数据源之间交叉引用信息。这对患者、入院或任何其他记录的不同技术和不同的识别方案增加了复杂性。此外，对软件数据库的直接写入访问权限通常不可用，因为编制软件人员很少授予此类权限，以防止对常规软件操作造成任何可能的中断。在事务性软件数据库中，数据由精心组织且高度准确的记录组成，这些记录以行的形式呈现。这些记录被非常精确地收集，以实现软件的特定功能。除数据外，还存在丰富的元数据，包括有关数据收集的信息（如有关输入数据的个人信息、记录时间戳和生物医学设备标识符）以及软件配置。值得注意的是，这些元数据的很大一部分可能与我们的研究目的没有直接关系，因为它们主要支持软件的日常功能。

数据湖

综合数据重用途径的第一个可选组件是数据湖。数据湖是一种集中、灵活且可扩展的数据存储系统，它以原始格式从多个异构源获取和存储原始数据。数据以细粒度、面向行和原始格式存储在安全且经济高效的环境中。这些原始数据仍然包含多种格式，从结构化数据到非结构化文本文档、图像、歌曲、视频和传感器数据，确保各种数据分析工作都能随时获得广泛的信息。

为数据湖实现的技术可以包括通常的关系数据库，如 PostgreSQL 或 Oracle，但也包括 NoSQL 数据库和大数据技术，如用于存储的 Hadoop 分布式文件系统或 Apache Hudi，以及用于数据处理的 Apache Spark、Hadoop MapReduce 或 Apache Kudu。

与通常集成到数据仓库中的结构化数据不同，数据湖不会立即进行结构化或转换，从而实现更敏捷、适应性更强的方法。这种灵活性使探索、操作以及必要时的数据转换成为可能，以满足特定的研究或分析要求。通过延迟预定义数据模型的应用，数据湖营造了一个环境，在这种环境中，无需预先确定的假设就可以发现信息。这包括在数据收集和存储的初始阶段可能不明显的见解。该系统进一步促进了即时查询处理 and 数据分析。

在没有数据湖的数据途径中，在利用数据之前，必须完成提取-转换-加载（ETL）过程。这引入了时间延迟，因为它需要识别 HIS 中的相关数据，更新数据仓库数据模型以适应这些数据，然后设计和实施 ETL。

此外，在解释结果时，如果分析明显缺少相关数据，则需要更新 ETL 过程和数据模型以合并缺失的数据。这种识别、修改和重新实施的迭代周期可能会导致时间线延长，并可能阻碍数据分析过程的敏捷性。因此，事实证明，数据湖法在为数据探索和分析提供更灵活和动态的环境方面具有优势，有可能避免在传统途径中遇到的一些挑战。

数据仓库

数据仓库是途径中最普遍的组件，充当来自 1 个或多个不同来源的集成数据的集中存储库。它以优化的格式存储历史和当前细粒度数据，以便进一步使用。这涉及单一存储技术、表和字段的一致命名约定以及跨数据源的一致标识符。这与数据湖不同，在数据湖中，所有这些元素在数据源之间都不同。

数据仓库是通过 ETL 过程。此过程的主要目标是从 HIS 或其他外部资源中选择和提取相关数据。在此初始阶段，与软件操作（如使用日志或界面设置）、监视器和输入数据的个人相关联的大多数元数据通常被排除在外。事实上，这些类型的元数据与患者护理信息无关，并且会引入不必要的数据量。随后，ETL 过程通过识别和纠正任何异常或错误信息来增强原始数据。在此优化之后，数据被集成到独立于源软件的统一数据模型中。值得注意的是，重点关注协调来自不同数据源的标识符，以确保数据完整性并简化涉及来自多个来源的信息的查询。ETL 过程还负责使用原始数据源中记录的新信息定期更新数据仓库。

数据仓库作为关系数据库，通常使用 PostgreSQL、Oracle、SQL Server、Apache Impala 或 Netezza 等系统实现。然而，对于数据仓库来说，探索 MongoDB、Cassandra 或 Couchbase 等 NoSQL 技术也很有趣，它们在处理非结构化或半结构化数据方面具有优势，并为大规模数据存储和检索提供可扩展性。可以使用 2 种类型的技术来开发 ETL 过程。第一种使用 R（R Core Team）、Python（Python 软件基础）或 Java（Oracle Corporation）等编程语言，可以与 Apache Airflow（Apache 软件基金会）等调度器结合使用，以组织角色执行以及日志和错误消息的检索。第二类应用是图形用户界面软件，如 Talend（Talend）或 Pentaho（Hitachi Vantara）。它们不需要编程能力，因为与数据管理操作相对应的图形组件是通过拖放界面进行组织的。

为了促进机构之间的合作并促进工具、方法和结果的共享，已经出现了一些提供通用数据模型（CDM）的举措。因此，表名和字段名按照通用命名法进行标准化，本地词汇表和术语映射到共享词汇表。在这些清洁发展机制中，观察性医疗结果伙伴关系 CDM 由观察性健康数据科学和信息学社区开发，该社区汇集了多个国家和数千名用户并导致了方法和实践上的进步。

因此，数据仓库对于细粒度的历史数据和元数据起到统一、集中和规范化的存储库的作用，并继续以面向行的格式呈现信息。Inmon 提出的建模方法，并被描述为“面向主题的、非易失性的、集成的、时变的数据集合”，这意味着数据是持久存储的，而不对其未来的使用进行任何假设，因此在使用中保持开放性。

数据市场

虽然数据仓库是一个独特的标准化存储库，主要用于数据存储，但由于关系模型中的数据量和分布，查询这些数据可能非常耗时。此外，集成到数据仓库中的原始数据可能不容易与特定的研究或分析问题保持一致，因为这些数据缺乏必要的聚合特征。例如，数据仓库保留所有生物学测量值（例如，钾和钠），而将存储在数据市场中的是与生物学值相关的特征，例如低钾血症、高钾血症、低钠血症或高钠血症的发生。因此，数据集市充当专用资源，用于将数据转换成可用且有意义的信息。此转换过程涉及特征提取，通过应用算法和特定领域的规则来实现。结果是解决特定研究问题或分析需求而量身定制的数据。例如，在临床环境中，数据市场可以将原始平均动脉压值转换为适合检测围手术期低血压的格式。

此外，数据市场可以以在线分析处理（OLAP）多维数据集的形式进行组织，从而提供数据的多维视图。这种立方体结构允许深入分析，使用户能够有效地探索和导航各个维度，例如时间、地理或特定类别，从而获得深刻和上下文化的见解。这些数据集市通常采用雪花或星形架构进行建模，从而优化其结构以创建 OLAP 多维数据集。星型架构（其中心事实表被维度表包涵）或雪花架构（进一步规范化维度表以最大程度地减少数据冗余）都有助于创建这些 OLAP 多维数据集。此类模式在提高 OLAP 环境中多维数据分析的效率方面发挥着关键作用，为更快、更全面的洞察提供了结构化框架。

在医疗保健方面，OLAP 立方体的一个例子可以包括患者（如年龄和性别）等维度、时间（如入院和出院日期）、医疗状况（如主要和次要诊断以及医疗程序）、医院单位（如有关服务、科室和床位类型的信息）、医务人员（如医生）和结局（如住院时间、治疗结局和医疗费用）。立体地包括各类事实，如患者人数、平均住院时间和平均治疗费用。这种多维结构使医疗保健专业人员能够进行深入分析，

探索随时间推移的趋势，比较不同医院单位的成本，并评估医疗干预对患者结果的影响。

由于数据的结构化性质，数据集市通常存在关系数据库（如 PostgreSQL、Oracle 和 SQL Server）中。对于 OLAP 多维数据集，这可能包括 Apache Kylin 或其他基于关系数据库构建的专有 OLAP 工具。

与 Inmon 的方法，金博尔自下而上的方法将数据集市置于核心位置，其设计主要由业务需求驱动。但是，通过直接开发数据集市，Kimball 方法可能会忽略一些在业务需求阶段最初未确定为相关的关键数据。

因此，数据集市是一个集中式组件，用于清理和聚合功能，用于专用目的，仍然存储在面向行的结构中。

功能商店

功能存储解决了数据市场中通常使用的传统面向行的关系数据库结构的局限性。此体系结构依赖于多个表，可能无法完全满足各种分析要求。例如，有效的统计分析通常需要一个具有面向列变量的平面文件，从而强制要求在特征存储中将数据从基于行的格式转换为基于列的格式。此过程简化了数据访问，将复杂的查询简化为从单个表中直接选择。因此，特征存储成为一个集中式存储库，其中包含有据可查、精心策划和访问控制的功能。除了从数据市场中提取的特征（通常由派生自业务规则的算法计算）之外，特征存储还可以接收机器学习算法生成的特征。

特征存储的设计旨在为数据科学家提供对这些特征的直接访问，无需进行额外的数据清理、聚合或透视。这一专业角色提高了效率并促进了高质量、可分析数据的使用，从而大大提高了医疗保健组织中数据驱动研究的有效性。值得注意的是，特征存储不仅存储特征本身，还存储其关联的元数据，记录如何计算和使用它们。它明确保留了所有特征版本，保证分析的可重复性。

当从业务规则派生时，功能存储在关系数据库（如 PostgreSQL、Oracle 和 SQL Server）或 NoSQL 数据存储（如 MongoDB）中，以存储元数据。当功能源自

机器学习模型时，它们将从 Databricks 或 Hopsworks 等大数据平台存储和共享。

作为数据重用途径的最后一个组件，特征存储在医疗保健组织内的各种分析应用程序中起着关键作用。它为创建富有洞察力的仪表板和自动报告做出了重大贡献，提供了实时和历史信息。在研究中，它最关键的贡献在于生成非规范化的平面表，类似于为统计分析量身定制的问卷数据。

结论

在这份研究中，我们提出了数据重用途径组件的标准化命名法和定义。表 2 总结了此途径中每个组件的优点和局限性。

虽然数据仓库是必要的初始阶段，但数据市场和特征存储的集成增强了其有效性。数据市场从原始数据中计算相关信息，而特征存储则将其组织到列中，从而简化数据集构建。此外，数据湖成为将原始数据存储在单个位置的宝贵资源，无需等待整个途径开发即可进行开发。

值得注意的是，在没有数据湖的数据途径中，在分析之前完成 ETL 过程的要求会带来延迟。这包括识别 HIS 中的相关数据、调整数据仓库数据模型以及实现 ETL。此外，在结果解释过程中发现缺失数据会促使对 ETL 流程和数据模型进行迭代更新，从而可能延长时间线并阻碍数据分析敏捷性。

需要强调的是，这里描述的特定组件及其特征不是一成不变的，可以根据独特的组织需求和配置而有所不同。例如，数据湖和特征存储的包含通常是自由裁量的，受源数据的规模和复杂性、特征的数量、研究项目的范围、团队的规模以及研究可重复性随时间推移的必要性等因素的影响。

表 2. 数据重用途径组件的优缺点

元件	优势	弱势
数据湖	同一服务器上的所有数据源 独立于源软件	数据格式和结构不一致 缺乏标准架构会使查询变得复杂

元件	优势	弱势
	实时查询处理 and 数据分析，无需完整开发提取-转换-加载（ETL）过程	分析重现性
数据仓库	统一的数据模型有助于查询来自行政系统和生物学系统的数据（即，来自两个系统的数据是链接的，并且模型约定是一致的）相关数据以最详细的程度保留（例如，日期、诊断和所有生物学值），从而能够回答许多问题，而无需事先确定它们	必须实施 ETL 过程以标准化数据 具有多个统计单元的多维数据模型 细粒度数据不能直接使用，也不能用于统计分析或决策
数据市场	功能已准备好直接使用	在多个数据市场中，特征仍以行格式（即每行 1 个特征）进行组织
特征存储	直接使用功能，无需执行数据管理任务，例如联接数据市场或透视以将功能重新组织到列中	事先开发了整个途径

译文二：

在法国地区大学医院中实施从蓝图到临床的生物医学数据仓库：揭示流程、克服挑战并提出临床见解

Matilde Karakachoff, Thomas Goronflot, Sandrine Coudol, Delphine Toubiant, Adrien Bazoge, Pacôme Constant Dit Beaufils, Emilie Varey, Christophe Leux, Nicolas Mauduit, Matthieu Wargny, Pierre-Antoine Gourraud, 徐健(译)

介绍

现在研究机构中越来越多地用到电子健康记录，为医生提供了从多个来源收集到的、系统的但能二次利用的数据。事实上，医院信息系统(HISs)面临着协调和集成应用系统和临床数据库的技术挑战，这些应用系统和临床数据库高度异构，基于编辑器特定的软件格式，并使用非标准化术语。

此外，由于国家和国际法律对患者隐私的规定，机构必须面对授予数据二次访问权的法律和伦理挑战。在护理过程中产生的数据的重用意味着必须通过定义良好的数据治理和访问策略来解决伦理和遗留概念的操作知识。存储库不仅要确保数据访问的技术方面，还要确保授予访问权限的决策标准。事实上，在考虑患者知情同意和隐私数据保护等法律和伦理原则时，机构数据治理也至关重要。最后但并非最不重要的是，按照传统流行病学的路线，临床数据重用需要在经过验证和标准化的方法框架内进行处理，以确保从科学和临床的角度获得定性结果。

尽管存在这些困难，生物医学数据仓库(BDWs)的兴起正在改变流行病学和临床研究的研究过程。患者数据构成了定义良好的概况，可用于促进队列的丰富、临床试验的患者选择和随访、表型检测和症状的详细描述。BDWs可以促进个性化和精准医疗的开发和实施，包括通过使用大数据和人工智能方法。

BDWs的实施在法国的各个地理层面上进行。据我们所知，2008年至2023年间，共有24家比较活跃的医院成立了BDWs。区域协调通常发生在BDW的专业网络中，例如“西部数据中心”，这是专门为法国西部的大学医院设计的。这可以发生

在专题网络中，并且在癌症数据中非常领先。国家或欧洲联盟范围内的倡议还提出了一个发展和协调框架，以应对执行方面的不同挑战。特别是，法国于 2019 年启动了一个名为“健康数据中心”的国家项目，旨在促进集中协调并提高全国范围内数据源的可见性。

本文报告了研究者在支持南特大学医院生物医学数据仓库 (NBDW) 实施中组织变革、技术架构和治理方面的 5 年经验。文章描述了访问临床内容的过程，还提供了有关存储库中包含的源和内容的图表，和对 NBDW 贡献项目的一些见解。最后，文章提出了三个指标来衡量设置操作的有效性。

方法

概述

2018 年，南特大学医院中心 (CHUN; 南特大学医院) 实施了 BDW，以促进最初在患者护理背景下收集的个人健康数据的二次利用，用于研究，并提供对 CHUN HIS 内不同来源的最新数据的单一安全访问，在一个统一的视图中包括人口统计和临床信息、咨询、计费代码、诊断、实验室结果、医疗记录和用药等广泛的数据类型。

旨在实施 NBDW 的干预重点需要协调相关医院研究实体的重组。这项重组举措不仅包括解决技术障碍，还包括划定治理结构、监管框架和管理数据访问的参数。

本实施报告遵循 iCHECK-DH (数字健康实施报告指南和清单) 报告准则。

从蓝图到 NBDW 的职能组织

NBDW 的实施需要在三层结构中重新组织一个功能研究网络 (图 1)。首先，治理政策和 NBDW 法律框架的执行是由研究行政部门 (RAD) 完成的，该部门承担了法律和道德责任。其次，临床数据中心 (CDC) 的建立负责科学问题和 NBDW 数据与项目研究者共同承担科学责任的潜在需求。最后一层技术责任是由医院的信息技术部 (ITD)

建立的。IT 优先级由法律和治理约束定义，并解决每个项目的数据需求。新的组织允许在 NBDW 的实施和管理中分配任务和责任。

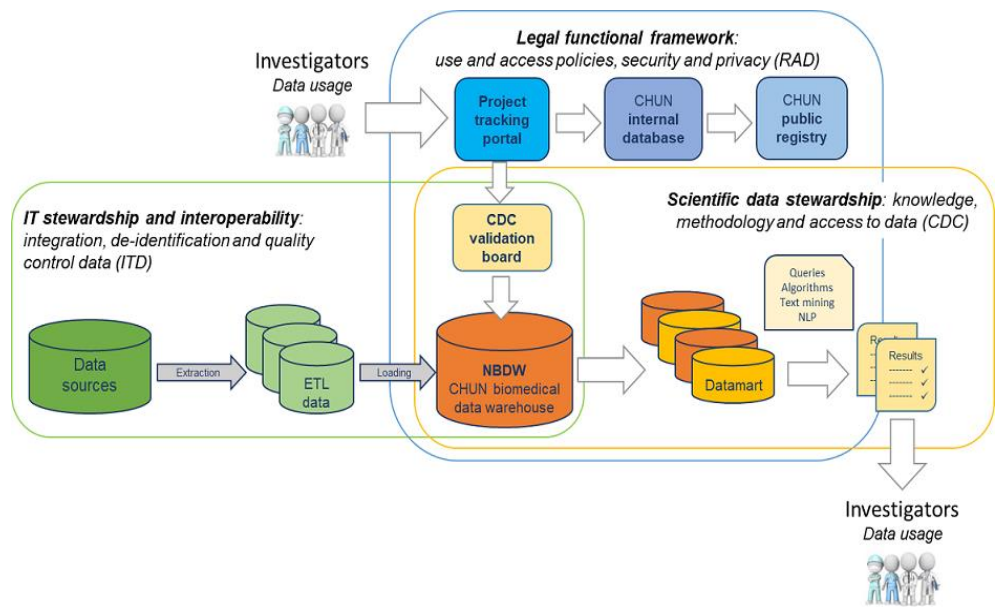


图 1 专用于 CHUN NBDW 的功能框架

三层治理结构构成了伴随 NBDW 发展的功能研究框架。ITD 确保 IT 基础设施的质量、透明度和数据去识别性。通过 ETL 过程，数据流被提取并加载到 NBDW 中。科技发展处是一个行政部门，负责法律和道德责任，并制定规管架构，界定资料使用和查阅政策。CDC 承担科学数据管理，负责科学责任，支持研究方法，并授予数据访问权。这三个结构是有组织和结构化的，以确保研究质量、伦理和技术透明度。白色箭头表示调查人员进行的数据请求和访问过程。CDC:临床数据中心;CHUN:南特大学医院中心;ETL:提取、转换和加载;ITD:资讯科技署;南特生物医学数据仓库;NLP:自然语言处理;RAD:研究行政部。

法律职能框架

RAD 代表医院定义数据使用和访问政策、安全性以及患者同意和隐私的框架，并负责 NBDW 监管框架的构建、组织和执行。它与医院数据保护官合作，由专门负责研究数据的官代表。此外，RAD 结构需要记录在案的监测，以确保 NBDW 监管框架不断调整，以应对法律标准的变化。所有项目都在网络上公开描述适用于所有提供数据的潜在患者。

信息技术管理和互操作性

ITD 负责临床数据的收集、存储、汇总、质量和完整性。该结构执行符合技术标准的连续过程。增加新的医院数据源是一个持续的过程，始于 2018 年 7 月，现仍具有重要意义。

NBDW 设置在虚拟机上。Oracle 数据库支持大规模数据存储，将多个医院数据源集成到一组独特的表中，其中包含相关患者、咨询、诊断、实验室结果、医疗记录和住院患者药物管理的数据。NBDW 采用 eHOP 软件对数据库进行组织和查询。eHOP 是雷恩大学在企业应用集成 Enovacom（奥兰治商业服务-桑达尔）的 Enovacom Suite Version 2 (ESV2) 上使用公私合作伙伴关系开发的平台。eHOP 对来自不同格式和标准的各种 HISs 的数据进行采集、转换和集成 (Health Level 7 [HL7]; 信息协调者和促进者 [HPRIM]; PN13; 逻辑观测标识名称和代码 [LOINC]; 以及 Word 文档、PDF、CSV 和文本)。ESV2 从 NBDW 提供自动的、预定的 (每日、每周或每月) 和监控的数据供应。

科学数据管理和对 BDW 的代理访问

从 2018 年开始，美国疾病预防控制中心（CDC）成立了一个专门的团队，致力于将健康数据再利用进行研究，依靠公共卫生医生、数据和计算机科学家、流行病学家、生物统计学家和项目经理的专业知识，促进流行病学分析，并为临床研究人员和研究人员提供数据访问方面的支持。CDC 与 ITD 一起定义了数据整合和数据使用实践所需的标准，并确保数据控制和分析的科学质量。

目标和数据访问：NBDW 的最终用户是旨在推进该机构研究的调查人员 (内部或外部的 CHUN)。通过由 4 个主要步骤组成的标准化审批流程，为 CHUN 调查员创建了每个项目的访问权限 (图 1)。首先，调查员通过由 CHUN 内部网托管的门户网站注册他们的请求并提交研究协议。其次，CDC 审定委员会每周召开一次会议，对提交的研究项目进行三个方面的分析：符合伦理原则、科学相关性和可行性。在完成审批过程后，该项目将在内部数据库中注册，以完成法律要求。第三，CDC 处理 NBDW 上所有必要的查询，以选择一组与科学问题相关的患者。在这一步骤中，与

研究者的持续合作是必要的，特别是要精确定义科学目的、符合条件的人群和感兴趣的数据。第四，数据是通过一个数据集市在内部提供的，这个数据集市由 CHUN 内部网托管。数据集市系统为调查人员提供规范的、简洁的（仅提供有关目标人群所需的数据）和有时间限制的访问。数据被完全识别。此外，研究人员可以通过 eHOP 进行简单的查询和查找数据集中包含的患者数据，eHOP 具有一组用于简单文本和结构化数据查询的工具。

需要数据管理和提取以整合研究数据库的项目向 CHUN 项目的公共登记处申报，以保证透明度并允许患者反对。在这个步骤中，可以应用更复杂的方法，通过自然语言处理(NLP)、正则表达式工具或其他结构化数据来提取信息。最后，数据提取被严格限制为必要的信息，遵循简约原则，并且只有在可以在安全环境中访问数据的情况下。

如果项目由来自 CHUN 的外部项目负责人(学术或私人合作伙伴)支持，则会进行上述相同的过程，但有以下不同:该项目可能由通过门户提交研究方案的临床团队支持;医院和合作伙伴之间必须签署合作协议，并且数据集市只能通过特定的虚拟工作空间使用(数据仍然在内部托管)。

数据保护和患者同意：为了遵守国家与国际隐私法规，数据集成受到去识别算法的约束。数据存储在两个独立的 Oracle 模式中，以将假名数据与指定或其他直接重新标识的信息分开，这些信息的访问受到严格限制。数据分离由访问管理和所执行操作的可追溯性(如审计)来补充。最值得注意的是，该平台包括收集和应用患者同意使用个人数据的功能，确保符合法国法律和欧洲通用数据保护条例要求。

监管许可：根据法国《数据保护法》（《信息与自由法》，1978 年），为健康研究和评价使用个人数据需要遵守代表良好做法的参考方法。如果没有这样的遵守，个人数据的使用必须得到国家信息和自由委员会(CNIL;法国国家信息技术和公民自由委员会)。在 NBDW 启动时，该领域还没有数据仓库的研究方法。因此，必须获得 CNIL 的批准才能启动实施。提交给 CNIL 的文件涵盖了法律责任、数据处理细节、访问、治理等内容。提供了全面的数据访问细节，扩展到研究人员，无论是否隶属于 CHUN。私营实体被允许从事基于 NBDW 的研究项目，确保遵守该决议和法国

法规。NBDW 的数据保护影响评估是提交给 CNIL 的一个组成部分，是一份强制性文件。CNIL 于 2018 年 7 月 19 日授权设立和使用 NBDW (第 2018 - 295 号决议)。

预算规划和可持续性：由于以下几个因素，估算与实现和维护数据仓库相关的成本具有挑战性。首先，NBDW 是机构战略的一部分，因此很难将其视为一个独立的实体。其次，实施它涉及多个机构和专家的协作和协调，使资源使用估计复杂化。第三，隐藏成本难以预测和考虑，包括系统故障和延迟、计划外的许可证续期和升级、监管和法律要求的调整、HISs 的计划外变更以及基础设施升级。我们考虑了三条预算线——基础设施、许可证和人力资源，以及两个不同的时期——5 年 (2018-2022 年) 完工，2 年 (2022-2023 年) 维护——总共 260 万欧元 (280 万美元)。

在可持续性方面，为维护和更新分派了年度业务预算。整合新医院数据源的具体需求是通过基于项目的资金提供的。此外，目前正在确定一种经济模式，以便在外部研究项目的情况下纳入基础设施费用的额外费用。

伦理声明：包含在 CNIL 第 2018 - 295 号决议的监管批准中。

实现(结果)

来源、概念和内容的说明

NBDW 将多个医院数据源集成到一组独特且结构化的表格中，其中包含有关患者人口统计和行政记录、住院患者药物管理、住院患者常数和人体测量分数和指标的数据、解剖病理学笔记、住院和门诊医学实验室结果、叙述性医疗笔记（包括入院/出院摘要、住院麻醉笔记、门诊咨询笔记、护士笔记），*国际疾病分类第 10 次修订版 (ICD-10)* 和法国医学分类 (CCAM; 医疗程序的通用分类) 住院诊断和程序的代码，以及医学影像报告。表 1 根据不同的 HIS 数据源或迄今为止集成的软件，显示主要概念和内容。某些数据源仅包含叙述性注释，而某些数据源同时包含非结构化和结构化数据。

表 1 南特大学医院生物医学数据仓库的主要来源和概念（数据提取于 2023 年 5 月 10 日）

概念	软件	时期	患者, n	记录, n	2022 年记录, n	结构化数据, n
住院药物处方	千年	2015-今	318, 456	39, 681, 513	5, 673, 000	248, 289, 146
心脏病学叙述性笔记	心脏报告	2015-今	9278	38, 041	6644	— ^a
会诊临床叙述笔记	GAM-CLINICOM 临床	2002-今	1, 053, 386	6, 507, 860	22, 982	—
常数和人体测量数	千年	2015-今	440, 250	3, 306, 589	638, 969	61, 142, 757
解剖病理学笔记	迪亚米奇	2015-今	131, 812	229, 081	27, 244	1246
生物学实验室结果	DXLAB 实验室	2012-今	701, 804	10, 752, 384	1, 672, 218	155, 825, 060
临床叙述笔记	千年	2015-今	569, 114	3, 967, 073	907, 573	5, 395, 233
ICD-10 型 ^b 和临床程序代码	临床	2006-今	725, 802	5, 318, 712	401, 607	105, 246, 620
放射学报告	QDOC 的	2015-今	284, 937	904, 625	122, 900	—
护士传输	传输	2017-今	131, 508	1, 546, 114	320, 293	1, 546, 114

^a 不适用

^b ICD-10: 国际疾病分类, 第 10 次修订本

NBDW 数字和人口

CHUN, 一家三级医院, 在活动方面在法国排名第七, 为 1 万居民的人口集水区提供护理。它为住院和门诊患者提供随访和长期医疗保健。它拥有 4 张病床, 接生 2993 名婴儿, 每年进行超过 4380 万次咨询和外部医疗程序。还为 1200 名医学生、800 名住院医师和 2000 多名非医学生开展实践教学。

NBDW 包括 1-5 年间约 2003 万名入院患者的信息（表 2）。超过 1 万例住院治疗与大约 2 万例 ICD-12 编码诊断和 3 万例临床程序代码相关。加上 10 多万次外部咨询, 《国家生物多样性公约》包含 7 3 多万份文本文件。这些叙述性笔记被整合为自由文本文档, 可以被询问并转化为结构化数据进行研究。

每年的患者人数、住院人数、咨询人数和叙述性笔记随着时间的推移而增加（多媒体附录 1），2003 年至 2019 年的增长率从 109%（住院）到 430%（门诊咨询）不等。

表 2. 南特大学医院生物医学数据仓库的数字和内容，2003–2022 年

内容	自 2003 年以来, n	限 2022 年, n
患者 ^a	1, 597, 498	300, 804
住院 ^b	2, 635, 809	183, 361
门诊咨询	6, 358, 271	524, 948
临床叙述笔记 ^c	11, 634, 761	826, 670
诊断 ^d	12, 251, 148	956, 688
临床程序 ^e	7, 272, 346	504, 571

a 具有≥1 临床叙述性笔记或结构化文件的患者，包括住院患者和因外出会诊而入院的患者；

b 内科、外科和产科服务的住院治疗，包括完全住院、日间住院和反复就诊；

c 临床叙述性笔记（不包括生命体征和人体测量数据；ICD-10 国际疾病分类，第 10 次修订版 [ICD-10] 和临床程序代码；实验室结果，；住院药物管理；护士传输不包括在内）；

d 根据 ICD-10 进行内科、外科和产科住院治疗的医学诊断；.

e 法国医疗程序通用分类 Classification Commune des Actes Médicaux - CCAM 用于内科、外科和产科住院。ICD-10：国际疾病分类，修订版。

项目和有效性成果

NBDW 数据的可用性使得有可能提供数据，以应对一系列的广泛科研问题和分析管理护理和组织的数据需求。在 NBDW 数据创建之前，研究人员仅限于数据重用许可范围之外的医疗管理结构化信息的查询。它只涵盖像与住院有关的 ICD-10 诊断的结构化数据、通过 CCAM 编码的医疗程序，较少是可用性问题和实验室结果。由于法律和技术原因，从不同 HISs 和数据管理器服务上独立执行的不同查询中获取结果是非常耗时的。

NBDW 目前在综合环境中以及数据保护政策和法律方面促进了结构化和非结构化自由文本注释中临床概念的查询。根据其目的，BDW 数据请求可分为三种不同类型的研究项目：（1）优化临床试验和观察性研究中的患者筛查，（2）丰富病例报告

或用于疾病监测的电子病例报告表，以及（3）评估和改进临床实践和资源管理。为了说明不同类型的研究，多媒体附录 2 中描述了三个具体的数据使用示例。

衡量 NBDW 有效性的第一个结果被定义为通过项目跟踪门户支持的研究数量（图 1）。自 2018 年以来，CDC 已提出并处理了 577 项请求。其中，269 个项目涉及由于 NBDW 查询和研究工具而纳入的患者（第二个结果），其中 115 个项目创建了数据集市，使研究人员能够访问数据（第三个结果）。

讨论

经验教训

临床数据仓库的发展为来自临床护理的大量不同数据提供了前所未有的访问。然而，它还需要在治理、与患者同意和数据保护相关的数据访问规则以及技术挑战方面付出一定努力。在功能研究框架内重组结构是 NBDW 成功的第一个因素。各部门之间的协作不仅促进了无缝沟通，而且还带来了处理医疗保健数据管理复杂性所需的创新。这是一个测试新 IT 技术（如分布式基础设施）和匿名化技术（如去标识化）的机会。在从 CNIL 获得 BDW 授权的过程中，结构之间的相互作用也是一个基本因素。事实上，获得监管部门的批准是长期谈判的结果，需要在满足法律、技术和科学研究要求方面具备专业知识（有关详细信息，请参阅“监管部门批准”部分）。

疾病预防控制中心的创建是由计算机科学家、NLP 工程师、统计学家、医生、流行病学家和项目经理组成的多学科团队合作的结果，可能是成功的第二个因素。在 RAD 和 ITD 结构的支持下，每周一次的 CDC 验证委员会从三个方面验证项目的合规性：伦理、科学相关性和实际可行性，在安全的背景下提供支持和对 NBDW 的访问。然而，它也是医院向数据驱动治理转变的主要驱动力。

NBDW 的实施需要一个持续且仍然相关的过程，以符合新的监管和技术标准，并增加新的医院来源和持续的改进。一个重要的教训是，在过去 269 年中，5 个 NBDW 项目中的每一个都是一个机会，可以重新审视 BDW 的内容，进一步推进质量控制过程，并领导数据转换过程，创造数据使用价值。

建立网络和共同努力可能是吸取的最好的教训。在法国，一些经验导致了卫生数

据预见的改变，并促进了区域间和国家中心的实施。NBDW 从“西部数据中心”中受益并做出了贡献，“西部数据中心”是一个由西法国大学医院数据仓库组成的网络。其目的既是促进数据分析的可重复性，共享资源和最佳查询做法，又是促进各中心之间适应和标准化的术语和命名法。此外，BDW 使用相同的软件进行集成和查询，从而允许使用一致的查询和规则来查询它们。这种方法确保了高水平的互操作性和可访问性，促进了无缝交互并遵守了可查找、可访问、可互操作、可重用（FAIR）原则。

事后看来，如果要重新审视这一进程，我们会以不同的方式处理某些方面。具体而言，在实施 NBDW 的初始阶段，主要重点是部署区域网络 ODH 建议的软件和 IT 基础设施，旨在促进存储库及其查询系统的建立。在承认这一战略固有的好处的同时，有必要对替代 IT 解决方案进行更彻底的检查，以减少对单一方法的依赖，并简化向替代和可变选项的潜在过渡。

结论

最后，由于缺乏系统的质控过程，使用电子健康记录进行健康研究需要特别注意，以确保结果准确。数据质量控制过程是一个漫长而必要的过程。未来的挑战将是建立标准化和共享的质量控制管道，以确保质量结果，不仅在地方一级，而且在未来数据共享的区域和国家背景下。通过扩大护理机构在 IT 和数据方面的长期投资，NLP 和文本挖掘工具的开发将进一步加速 BDW 的使用，促进从高层管理人员到患者的数据驱动决策讨论。

在卫生管理机构开展的任何研究项目和基于护理的研究分析都可受益于部署有组织的数据访问。数据生产的协作性质以及对患者的信息和隐私保护需要通过代理和专家访问 BDW。它证明了技术解决方案是更好的数据驱动实践的部分答案，并且必须导致明确的治理策略。

***注：原文和译文版权分属作者和译者所有，若转载、引用或发表，请标明出处。**