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 2023.8.25

（七）数据仓库

1. **标题：调整健康记录的逐步数字化：多医院临床数据仓库的工作示例**

来源： medRxiv preprint.

时间： 2023 年 8 月.

链接： doi: https://doi.org/10.1101/2023.08.17.23294220.

概要：医疗保健持续数字化产生了越来越多富含患者信息的数据，为研究、创新、质量监测和公共卫生监测创造了机遇，但同时给数据分析提出了新挑战，比如，临床数据仓库（CDW）包括越来越多的医疗站点，每个站点可能都有自己的特征，需要分析的数据也越来越多样化，包括诊断代码、临床报告、实验室测试、成像、药物等，所以，数字化快速发展导致数据可用性、格式和覆盖范围的频繁变化，阻碍了多中心纵向研究的进行，亟待引入创新方法解决这一难题。本研究为解决大规模临床数据分析中因健康记录日益数字化而导致的时间依赖性数据缺失问题提出了一种新的数据驱动分析法，并用此方法对大巴黎大学医院临床数据仓库收集的 38 家医院从 2013 年至 2022 年逐步采用通用电子健康记录的情况进行了分析。研究是根据数据的类别在医院、部门或单位分别展开，先通过回顾性队列研究评估了分析模型的性能，再通过数据驱动的两种方法（单纯分析法和纯源分析法）测量了质量和流行病学指标的时间变化。研究结果表明，采用单纯分析方法时，质量和流行病学指标存在不切实际的时间变化，而采用纯源分析法时，这种影响大大降低或消失。这说明数据驱动法可以用来解释由健康记录的逐步数字化引起的缺失；但是纯源分析法并不是万能的，需要在研究时间跨度上找到一个最优权衡；还有，本研究的重点是住院、急诊科和重症监护病房的记录，以及诊断代码、出院处方和会诊报告，所以其他数据类别可能需要对其关联的数据源进行特定的建模，未在本研究范围内。最后得出研究结论，电子病历研究需要使用不断发展的系统收集的数据。数字化的多尺度建模可用于减轻对医疗质量和流行病学研究中的虚假时间变化。

2. **标题：临床数据仓库实施的良好实践：法国案例研究**

来源： PLOS Digit Health.

时间： 2023 年 7 月.

链接： https://doi.org/10.1371/journal.pdig.0000298.

概要： 真实数据(RWD)在提高医疗质量方面有着巨大的前景，但是需要用特定的基础设施和方法来获得可靠的知识并为患者带来创新。根据对法国地区和大学的 32 家医院治理的国家案例研究，我们看到现代临床数据仓库(CDW)的关键影响因素:治理、透明度、数据类型、数据重用、技术工具、文档和数据质量控制流程。研究在 2022 年 3 月-11 月间以半结构化的方式进行了半结构化访谈，并对法国 CDW 报告的研究进行了回顾。在法国的 32 家地区和大学医院中，14 家正在产生 CDW，5 家正在试验，5 家有一个潜在的 CDW 项目，8 家在撰写本文时没有任何 CDW 项目。法国从 2011 年开始实施 CDW，并在 2020 年底开始加速实施。本研究总结这些案例研究，为 CDW 建立一些一般性指导方针。CDW 在面向实际研究时需要在治理稳定、数据模式标准化以及数据质量和数据文档的开发等方面做出努力，还必须特别注意的是仓库团队的可持续性和多层治理、研究的透明度和数据转换工具的提升、允许成功的多中心数据重用以及常规护理的创新。

常规护理数据的重复使用并非免费。为创造稳健的知识和创新发展，在应用时必须关注数据的整个生命周期。在对法国 CDW 进行首次概述的基础上，我们将常规护理数据收集和组织的关键方面记录到同质数据库中：治理、透明度、数据类型、数据重用主要目标、技术工具、文档和数据质量控制流程。法国 CDW 的发展呈现出一种渐进但仍不完全的同质化趋势。世界一些国家和欧洲的项目正呈现出支持标准化、方法学工作和工具方面的本地化。基于上述 CDW 样本分析，本研究得出了旨在巩固常规护理数据潜力以改善医疗保健的一些建议，即：创建和维护能够操作 CDW 并支持各种项目的多学科仓库团队；建立一个多层次的协作网络；鼓励建立一种通用数据模型，使用精确的元数据来映射集成的数据；将数据应用扩至医院领域之外。

（徐健编辑）

译文一：

调整健康记录的逐步数字化：多医院临床数据仓库的工作示例

View ORCID ProfileAdam Remaki, Benoît Playe, Paul Bernard, Simon Vittoz, View ORCID ProfileMatthieu Doutreligne, Gilles Chatelier, View ORCID ProfileEtienne Audureau, View ORCID ProfileEmmanuelle Kempf, View ORCID ProfileRaphaël Porcher, View ORCID ProfileRomain Bey,徐健（译）

1 背景和意义

医疗保健的持续数字化正在产生越来越多的常规数据。这些数据包含患者的丰富信息，为研究、创新、质量监测和公共卫生监测开辟了新的前景。然而，对他们的分析也提出了新的方法挑战，这些挑战在许多方面是与专门以研究为目的而收集的数据相关。特别是，数字化的快速发展导致的数据可用性、格式和覆盖范围的频繁变化，从而阻碍了多中心纵向研究的进行，因为在不同背景下收集的记录并不总是具有可比性。为了解决这个普遍存在的问题，一些研究提出了有助于检测时间推移或统计分布突变的工具，而另一些研究则是探索了在分析之前使数据集同质化的方法，例如插补缺失数据或删除不完整的案例。这些开创性的研究为常规数据分析铺平了道路，但仍有许多问题有待解决。

首先，统计分布的时间变化可能是由技术变革引起的，也可能是由其他机制引起的。将技术引起的变化与由护理重组或患者群体变化引起的变化分开似乎是回答许多感兴趣问题的先决条件。

其次，需要考虑的技术变革越来越复杂。例如，临床数据仓库（CDW）包括越来越多的医疗保健站点，在这些站点中，每个部门可能都有自己的特征。此外，正在分析的数据越来越多样化，包括诊断代码、临床报告、实验室测试、成像、药物等，从而增加了要考虑的动态数量。因此，仅通过目测来控制这些复杂的数据分布变化将要调动大量资源，并且显然需要一定程度的自动化。

2 目标

在这项研究中，我们开发并评估了一种自动处理医院健康记录逐步数字化引起的时间依赖性缺失的新方法。我们分析了在大巴黎大学医院（援助公共医院，APHP）收集的数据。在可能导致数据丢失的众多机制中，我们关注的是逐步采用用于收集行政记录、诊断代码、出院处方和咨询报告的功能的效果。使用数据驱动的方法，我们模拟了不同级别（医院、部门、单位）采用它们的步伐。然后，我们通过测量一些质量和流行病学指标来评估该详细模型的性能。我们使用两种不同的方法评估了它们随时间推移的稳定性，一种是没有考虑到健康记录逐渐数字化的简单方法，另一种是利用前建模的纯源分析的新方法。这项工作旨在回答以下问题：

- 我们可否利用纯数据法自动对健康记录的渐进数字化进行建模？
- 在进行原型观察研究时，我们可否使用这样的模型来调整这种机制引起的时间依赖性缺失？

3 材料和方法

该研究由 AP-HP 的机构审查委员会审查和批准（IRB00011591，决定 CSE21-33）。法国法规不要求此类研究的患者书面同意。根据《欧洲通用数据保护条例》，患者被告知，那些反对将其数据二次用于研究的人被排除在研究之外。本报告遵循补充材料 F 部分中使用观察性常规收集的健康数据（RECORD）报告指南进行的研究的重新移植。

3.1 数据源

AP-HP 包括遍布巴黎地区的 38 家医院(2.2 万张床位，每年 150 万人次住院)。自 2012 年以来，通用的电子健康记录(EHR)软件 ORBIS Dedalus Healthcare 已逐步采用。本研究中考虑的六类数据中每一类都使用不同的 EHR 功能收集(即与住院有关的行政记录、急诊科(ED)或重症监护病房(ICU)、诊断代码、出院处方和咨询报告)。CDW 遵循观察性医疗结果伙伴关系-通用数据模型 5.4 版标准。数据每天被整合到 CDW 中。研究从 2023 年 7 月 31 日开始。

3.2 EHR 采用的建模

AP-HP 通用 EHR 的应用是通过功能实现的，每类数据的收集取决于特定功能的使用。专用于住院记录的功能在医院级别采用，而其他功能通常在部门或单位采用（如急诊室或 ICU 就诊、诊断代码或临床报告）。目前没有精选知识库提供有关有

效使用每个 EHR 功能的信息，因此我们采用了数据驱动的方法来生成。由于不同机制的相互作用，从数据中提取 EHR 采用的动态并不简单：i）EHR 功能的技术引入导致数据可用性突然增加，但 ii）临床医生对这些功能的使用又会使这种效果趋于缓和，有时也会被 iii)将早期软件中收集的数据复制到新的电子病历中。这种机制会影响数据可用性曲线的形状，随着时间的推移，数据可用性曲线很少遵循理想的步进形状。然而，与其他机制(如临床实践的变化)引起的变化相比，引入和采用新的电子病历功能通常会导致数据可用性的突然变化，这些变化在时间上和在医疗保健站点内更加本地化。因此，为了自动检测 EHR 的采用情况，我们计算了 c（t），即每个 EHR 功能和每个医疗保健站点（即医院、部门或单位）在 t 月份的完整性估计值，并将阶跃函数拟合到这些突然变化的时间序列中（有关详细信息请参阅补充材料的第 A.1 节）。

根据 EHR 功能，我们采用了 c（t）的两个定义：具有至少一个数据点的住院记录的比例（用于研究用于收集诊断代码和出院处方的 EHR 功能），或者，当没有这样的分母时，每月数据点的数量除以研究期间测量的最高值（研究住院记录、急诊记录、重症监护室住院访问和咨询报告）。此建模为每个医疗保健站点和每个 EHR 功能提供了 t_0、预计收养日期和 c_0，即该日期之后完整性的稳定平均值。在访问 ICU 的情况下，我们还评估了一种替代建模，该模型包括将矩形函数拟合到完整性估计而不是阶梯函数，从而额外提供了 t_1，对应于单元中数据收集结束的灭绝日期。这种替代建模的动机是观察到，由于医院重组，在单个 ICU 中收集的数据有时会在某个日期后消失（见补充中的图 S1）。最后，为了评估模型的拟合优度，我们计算了一个*误差*项，该误差项由 c 之间的均方误差定义。$_0$和 t 之后*的* c（t）$_0$（见补编中的图 S2）。

3.3 质量和流行病学指标

EHR 数据可用于回顾性研究或监测前瞻性质量或流行病学指标。然而，随时间变化的缺失可能导致对这些指标的估计有偏差。在本文中，质量指标被定义为观察到某些结局的每月住院比例，流行病学指标被定义为与某些季节性流行病相关的每周住院人数。表 1 列出了我们考虑的指标以及我们用于选择每个队列和计算结果的数据类别。使用 J 21 和 J 09，J 10，J11 选择细支气管相关和流感相关住院治疗国际疾病分类 10 th修订版（ICD-10）代码。

Indicator	Definition	EHR functionality used for cohort selection (adoption level)	EHR functionalities used for outcome measurement (adoption level)
30-day rehospitalization	Proportion of hospitalizations with a rehospitalization in the 30 days after discharge.[32, 33]	Hospitalization record (hospital-level)	Hospitalization record (hospital-level)
30-day ED consultation*	Proportion of hospitalizations with at least one consultation in ED in the 30 days after discharge.	Hospitalization record (hospital-level)	Emergency record (hospital-level)
30-day consultation	Proportion of hospitalizations with at least one outpatient visit occurring in the 30 days after discharge.[34]	Hospitalization record (hospital-level)	Consultation reports (department-level)
Discharge prescription	Proportion of hospitalizations with at least one discharge prescription delivered to the patient.	Hospitalization record (hospital-level)	Discharge prescription (department-level)
30-day ICU readmission*	Proportion of hospitalizations with at least one readmission in an ICU in the 30 days after discharge.[35]	Hospitalization record (hospital-level)	Intra-hospitalization visit to ICU* (unit-level)
Bronchiolitis-related hospitalizations	Weekly number of hospitalizations with a bronchiolitis diagnosis.[36]	Hospitalization record (hospital-level)	Diagnostic code (department-level)
Flu-related hospitalizations	Weekly number of hospitalizations with a flu diagnosis	Hospitalization record (hospital-level)	Diagnostic code (department-level)

*ICU and ED stand for intensive care units and emergency departments, respectively.

表 1：质量指标和流行病学指标

3.4 统计分析

连续变量报告为四分位数范围的中位数(IQRs)，定性变量报告为数字和比例(%)。计算了从开始日期可变的 t_{init} 到结束日期固定的 2022 年 5 月期间的质量和流行病学指标。我们之所以选择这个结束日期，是因为一些技术问题影响了在此日期之后将临床报告整合到 CDW 中。质量指标的时间变化采用线性函数建模:

$$QI(t) = \alpha_0 + \alpha_1 t + \epsilon(t) \quad (1)$$

以 QI 为质量指标，t 为月，α_0 和 α_1 参数分别表征原点和线性趋势，*以及* ε（t） 随机误差。我们使用普通最小二乘回归估计模型系数和 95%置信区间（CI）。我们讨论了线性趋势 α_1 因为它表征了时间变化，并且可能受到随时间变化的数据缺失的影响。定性讨论了流行病学指标的时间变化，特别强调了 COVID-19 后时期。事实上，细支气管炎和流感的季节性流行受到 COVID-19 的影响，流行病学指标被用来调整医疗机构的反应。

我们使用单纯分析方法（N）或一种新方法（我们称之为纯源分析法（CSO））进行这些分析，该方法没有考虑到医疗保健的逐步数字化。为了计算结果，我们要么使用所有可用数据（N），要么将分析限制在医疗保健站点（医院、部门或单位），其中用于收集所需数据的 EHR 功能在研究期开始之前被认为完全采用（CSO 方法，$t_0 \leqslant t_{init}$ 对于每个医疗保健站点和用于计算结果的每个数据类别，如表 1 所述）。为简便起见，考虑到自 2013 年 1 月以来收集住院记录的 38 家医院中的 28

家医院的住院队列(即在该日期之前的住院记录为 0)，两种方法使用相同的分母计算质量指标。

我们期望在使用单纯分析方法时观察到数值递增的指标，因为 EHR 功能的逐步采用会导致对结果检测的时间改善。相反，CSO 方法旨在稳定用于检测结果的数据源，以避免这种虚假的时间漂移。

我们进行了两次敏感性分析来检验质量指标。首先，我们在 2013 年 1 月、2016 年 1 月和 2019 年 1 月改变了 t_{init} 的值。其次，我们对每家医院进行了亚组分析。统计分析使用 Python package statmodels 进行。EHR 采用的建模是使用免费提供的 Python 库 EDS-TeVa v0.2.4 实现的。

4 结果

4.1 EHR 采用的建模

APHP CDW 包含了 14 万患者相关数据。CDW 中可用的数据总量随着时间的推移而增加，其动态因数据类别而异，并反映了医疗保健的逐步数字化（见图 5）。虽然与住院有关的行政记录显示了十年来的收集稳定，但其他数据类别显示收集的单调增加，反映了新的 EHR 功能的逐步采用。

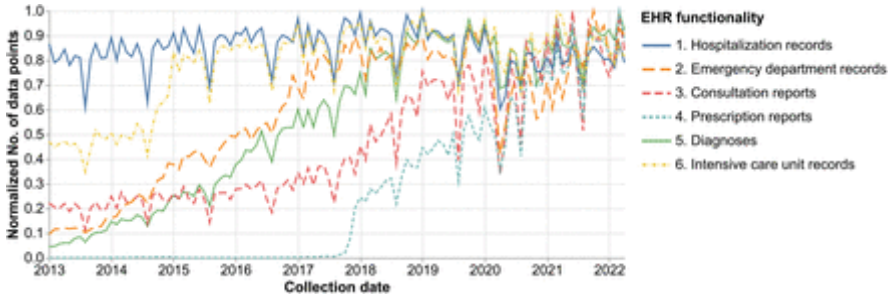

图 1：每个功能在电子健康记录中收集的数据量的时间演变。为了获得通用量表，对于每个功能，每月数据点数除以研究期间测量的最高值。

如图 2 所示，我们通过拟合步骤功能与适合采用机制（即医院、部门或单位）的描述级别来模拟每个医疗保健站点中这些功能的采用。住院、急诊室和 ICU 记录的数字化是突然的，而采用 EHR 功能来捕获处方报告或咨询报告则更加缓慢。虽然每个医院和每个科室的数据可用性曲线大多是阶梯形的，但当我们查看较小的单位级别时，在这里查看 ICU 记录，我们观察到采用是矩形的（见图 S1）。

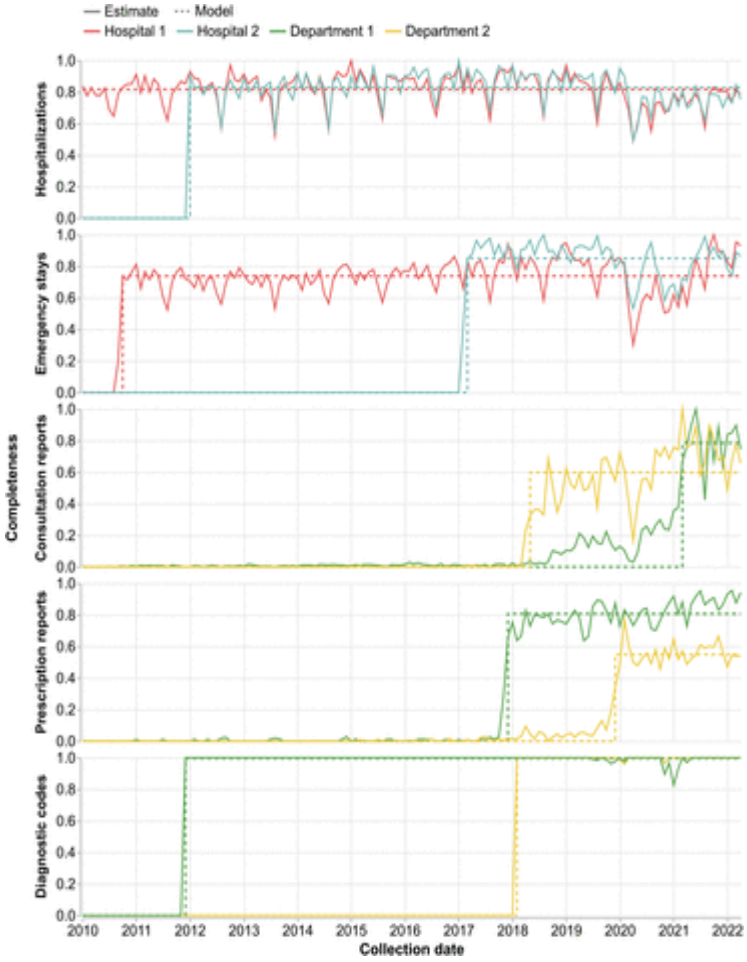

图 2：任意选择的医院和科室的完整性估计（平线），考虑从上到下的住院记录、急诊科记录、咨询报告、出院处方和诊断代码。1 号和 2 号病房分别位于 1 号医院和 2 号医院内。完整性估计定义为每月数据点数除以其在研究期间的最大值（住院和急诊科记录以及咨询报告），或定义为有至少一个数据点的住院记录的比例（诊断代码和处方报告）。完整性估计的建模显示为虚线曲线。

4.2 质量指标

图 3 显示了使用单纯方法或完全源方法的质量指标的时间变化，并选择了三个不同的开始日期。两种方法得出的 30 天再住院曲线相似，但其他指标存在差异。对于 30 天 ED 会诊、30 天会诊和出院处方，采用朴素法观察到单调增加，采用 CSO 法则消失。然而，CSO 方法的应用降低了指标的绝对值，因为用于确定结果的数据经过过滤以获得数据源的时间稳定性。在较近的起始日期，这种减少幅度较小。采用朴素法时，ICU 再入院率也呈现单调增加，而采用步长函数 CSO 法时，该指标呈现单调下降(图 S3)。使用矩形函数 CSO，也考虑到单位的消失，产生了一个更稳定的指标，正如预期的那样。

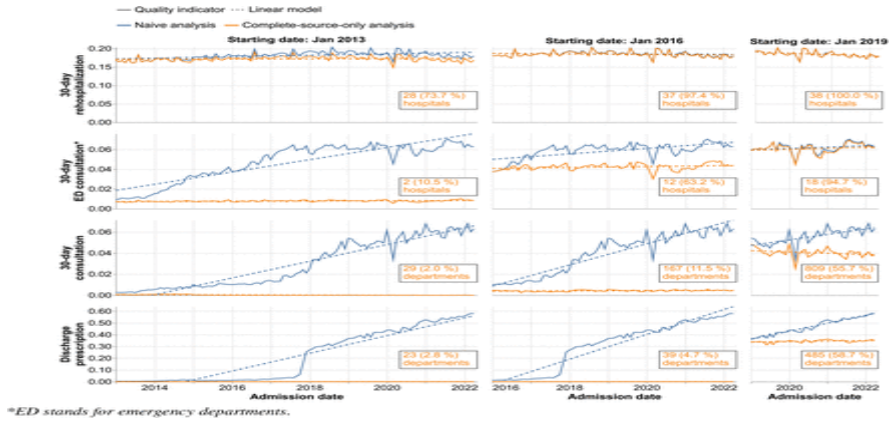

图 3：改变初始观测日期(tinit)和分析(蓝色为幼稚，橙色为完全来源)对质量指标纵向研究的综合影响。时间变化的线性模型用虚线表示。插图显示了在完全来源分析中选择的医疗保健站点的数量，以及它们所代表的医疗保健站点的比例，这些站点在研究期间(2013 年 1 月至 2022 年 5 月)至少有一个数据点由 EHR 功能收集。

如敏感性分析（图 4）所示，当单独考虑用于队列选择的 28 家医院时，仍然观察到 CSO 方法的稳定效果。该灵敏度分析还显示，CSO 方法诱导的指标振幅大幅降低，与图 3 中的观察结果一致。对于 30 天 ICU 再入院，对 28 家医院的敏感性分析显示，较旧的基线数据的朴素分析呈正斜率，阶梯函数 CSO 方法的斜率为负斜率，矩形函数 CSO 方法没有斜率（图 S4）。与其他质量指标类似，敏感性分析表明，两种 CSO 方法都降低了指标评估的幅度。所有拟合参数均可在补充材料的表 S1 至 S6 中找到。

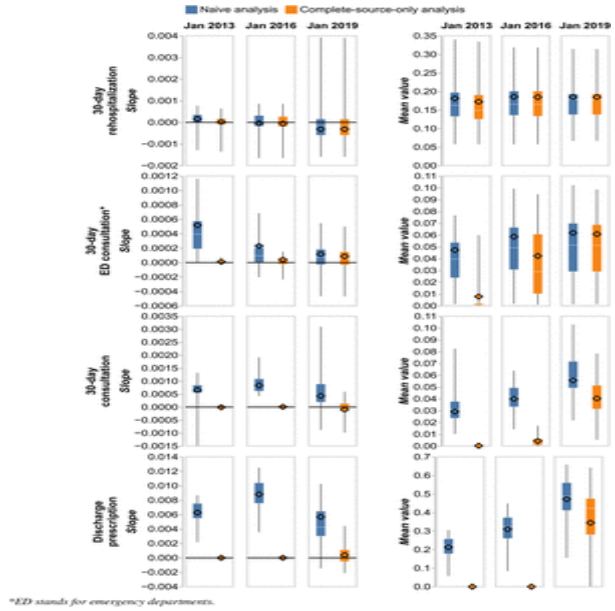

图 4：左：斜坡 α_1 线性模型（方程 1）估计在所有医院（菱形）或单独考虑每家医院（IQR 是盒子，最小和最大分别是下胡须和上胡须），采用朴素（蓝色）或仅完整源（橙色）方法。考虑了三个不同的初始日期。右：研究期间平均的指标值。

4.3 流行病学指标

图 5 显示了使用在 EHR（N）中收集的所有诊断代码或仅在自观察开始（CSO）以来完全采用 EHR 功能的部门收集的细支气管炎和流感相关住院人数估计的每周住院人数。虽然在这两种情况下都观察到 COVID-19 背景对这些流行病的强烈影响，但在使用朴素方法时似乎很难解释曲线，因为 COVID-19 效应无法与逐步采用 EHR 的影响分开。事实上，从 COVID-19 爆发之前的几年可以看出，这种机制导致了季节性流行病幅度的虚假增加。

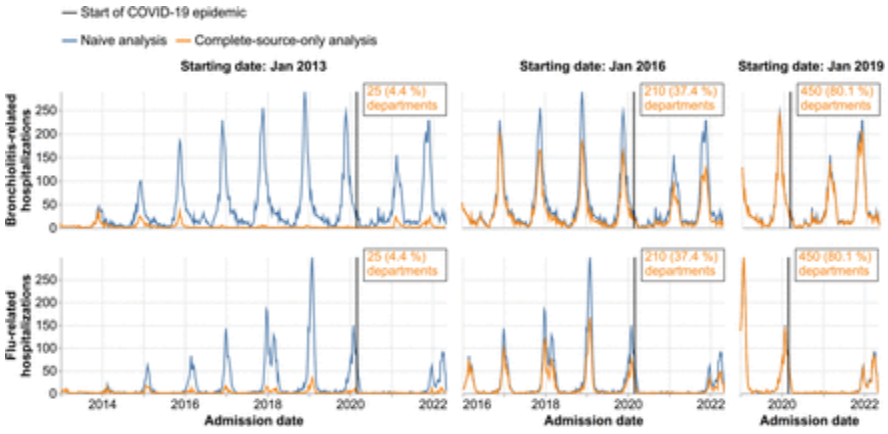

图 5：采用原始(蓝色)或完全来源(橙色)方法计算流行病学指标，并改变初始观察日期(tinit)。灰线表示 COVID-19 流行的开始时间(2020 年 3 月)。插图显示了在完全来源分析中选择的医疗保健站点的数量，以及它们所代表的医疗保健站点的比例，这些站点在研究期间(2013 年 1 月至 2022 年 5 月)至少有一个数据点由 EHR 功能收集。

5 讨论

我们表明，数字化的多尺度建模可用于减轻质量和流行病学指标研究中的虚假时间变化。为此，我们放弃了自测量期开始以来尚未完全采用所需 EHR 功能的所有医疗保健站点，并将此方法称为纯源方法。

我们在做任何修正之前都观察到完全随时间的变化与之前工作一致。我们的调整方法接近于众所周知的完整病例方法，主要区别在于我们采用了每个医疗站点的方法，而不是每个病例的方法，以便利用我们对渐进式数字化引起的缺失机制的理解。通过利用采用这种描述时出现的特定模式，即数据可用性的突然增加或减少，我们能够专门调整由逐步采用 EHR 功能引起的时间依赖性偏差，从而补充其他专注于数据分布的总体漂移或转移的工作。正如 Finlayson 等人所讨论的那样，许多机制可能导致时间变化，我们强调这项工作只解决了其中一个问题。

-- / --

CSO 方法不是万灵药。一方面，它的影响很大程度上取决于所研究的时间跨度。虽然时间太长会导致大量的数据丢失，但时间太短会严重限制纵向研究的科学相关性。因此，应该找到一个最优的权衡。此外，这种方法只有在统计分布的时间漂移不利于研究目标时才有用。例如，如果目标是在承受时间依赖偏差的情况下最大化检测到的事件/结果/患者数量，则不合适。另一方面，省略数据来源并非总是无害的，因为它可能通过改变所研究的人群而导致选择偏差，例如，省略患有不同严重程度患者的医院单位。因此，对使用这种策略获得的结果的解释应保持谨慎。

除了 CSO 方法的特定优势和局限性之外，本研究还说明了与现实世界数据平台相关的一些组织挑战。出于隐私考虑，调查人员只能访问从包含数百万条记录的完整数据库中提取的最小化患者队列。然而，本研究中提出的分析工作流的应用需要背景信息，如对整个数据库的完整性估计。因此，除了患者层面的数据外，这些指标应预先计算并提供给研究人员，这需要平台操作员和研究团队之间的密切协调。为了应对这一挑战，我们将该项目的计算机代码构建为一个开源库，研究人员可以对其进行扩展，同时平台操作员可以将其应用于整个数据库(参见图 5 所示的工作流程)。

我们的研究有几个局限性。首先，我们使用了单一的、高度简化的完整性定义。在某些情况下，可以对其进行改进，以匹配更直观的定义或更详细的合理性定义。其次，我们使用了 EHR 采用的粗略模型，该模型依赖于对数据可用性突然变化的检测。医院信息系统的实际动态表现出不同的行为，可能会在统计分布中引起更复杂的模式。第三，我们考虑了高度简化的质量和流行病学指标。因此，我们计算的指标并不直接适用于流行病学监测或质量监测。特别是，进一步描述分析中包括的患者特征似乎很重要，以便能够比较具有类似病例组合的设施。第四，我们的分析侧重于一些行政和临床数据类别，这些类别没有涵盖 CDW 中发现的各种数据。因此，应扩大它的范围，以支持对真实世界数据进行的更大范围的研究。

6 结论

电子病历研究需要使用不断发展的系统收集的数据。目前技术创新的步伐不太可能在短期内放缓，而且这种复杂性将继续增加。通过关注特定的机制，采用新的 EHR 功能，我们已经展示了各种元数据可以用于有意义地大规模分析 EHR 数据。此

外，自动化此类元数据的计算对于避免数据管理负担激增至关重要。我们的工作是朝着开发应对这些挑战的工具和方法的方向迈出的一步。仍有许多工作要做，特别是将有关技术变化的资料充分纳入研究的统计设计。

译文二：

临床数据仓库实施的良好实践：法国案例研究

Matthieu Doutreligne, Adeline Degremont, Pierre-Alain Jachiet,
Antoine Lamer, Xavier Tannier,徐健（译）

1. 介绍

1.1 真实世界的数据

HIS 越来越多地收集常规护理数据。这种真实世界数据来源（RWD）有望提高护理质量。一方面，这些数据的使用是开发个性化医疗的基石，从而转化为患者的直接利益（主要用途）。它们还通过加速和改善知识生产带来间接益处——二次利用：病理学、卫生产品和技术的使用条件、安全性、日常实践中的功效或有用性。它们还可用于评估卫生产品和技术的组织影响。

近年来，许多国家的卫生机构开展了大量工作，以更好地支持真实数据的生成和使用。监管机构也已启动了研究项目：欧洲药品管理局的 DARWIN 欧盟项目和美国食品和药物管理局的真实世界证据项目。

1.2 临床数据仓库（CDW）

在实践中，动员这些常规收集的数据的可能性在很大程度上取决于它们的集中程度，梯度从单个同质 HIS 的集中到具有异构格式的众多 HIS 的碎片化。HIS 的结构反映了治理结构。因此，处理这些数据的难易程度在很大程度上取决于医疗保健参与者的组织。RWD 的两个主要来源是保险索赔（更集中）和临床数据（更分散）。

索赔数据通常由国家机构收集到集中存储库中。在韩国，负责医疗保健系统性能和质量（HIRA）的政府机构与所有医疗保健利益相关者的 HIS 相关联。HIRA 数据包括国家保险索赔。英格兰在国家卫生服务（NHS）下有一个集中的医疗保健系统。尽管没有详细的临床数据，但这允许 NHS 将索赔数据与来自 2 个大型城市医学数据库的详细数据合并，对应于 2 个主要软件出版商。这些数据目前通过 Opensafely 访问，Opensafe 是第一个专注于 2019 年冠状病毒病（COVID-19）研究的平台。在美国，即使分散在不同的保险提供商之间，索赔也会汇集到大型数据库

中，例如 Medicare，Medicaid 或 IBM MarketScan。最后，在德国，不同的联邦主张直到最近才集中起来。

另一方面，**临床数据**往往分布在许多实体之间，这些实体做出了不同的选择，没有共同的管理或互操作性。但大型机构数据共享网络开始出现。韩国最近发起一项倡议，建立一个专注于重症监护的全国性数据网络。美国正在构建 Chorus4ai，这是一个汇集来自 14 所大学医院的数据的分析平台。为了释放临床数据的潜力，德国医学信息学计划在 4 年内建立 2018 个联盟。他们旨在开发技术和组织解决方案，以提高临床数据的一致性。

以色列是少数几个大规模汇集索赔和临床数据的国家之一，其一半人口依赖于一个单一的医疗保健提供者和保险公司。

基础设施将来自一个或多个医疗信息系统的数据（无论组织框架如何）汇集为同构格式，以便管理、研究或护理重用图 1 说明了 CDW 的 4 个阶段，数据流来自构成 HIS 的各种来源：

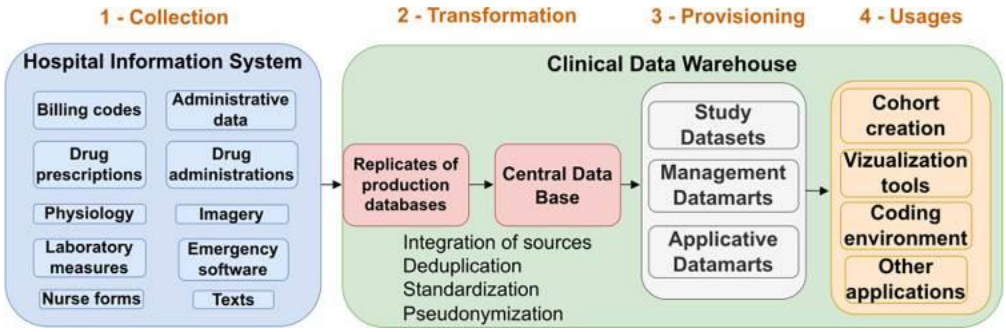

图 1 CDW：来自医院信息系统的数据流的四个步骤：（1） 收集、（2） 转换和 （3） 配置。CDW，临床数据仓库

1. 收集和复制原始资料。
2. 转型：整合与协调。
    - 将源集成到一个独特的数据库中。
    - 删除标识符重复数据。
    - 标准化：独立于软件模型的独特数据模型协调了通用模式中的不同来源，可能具有通用命名法。
    - 假名化：删除可直接识别的元素。
3. 提供子群数据集和转换后的数据集，用于主要和次要重用。
4. 用法要感谢专用的应用程序和工具来访问数据集和数据集。

在法国，国家保险公司将所有医院活动和城市护理索赔收集到一个独特的报销数据库中。然而，临床数据历来分散在众多 HIS 的每个护理站点。几家医院花了大约 10 年的时间努力从电子病历创建 CDW。随着 CDW 开始在区域和国家层面构建，这项工作最近正在加快。区域合作网络正在建立，如西部数据中心。2022 年 7 月，卫生部发起了一项 5000 万欧元的项目呼吁，旨在到 2025 年建立和加强与国家平台卫生数据中心协调的医院 CDWs 网络。

1.3 目的

基于对法国大学医院 CDW 的概述，本研究提出了正确利用 CDW 的潜力来改善医疗保健的一般建议。它侧重于：治理、透明度、数据类型、数据重用、技术工具、文档和数据质量控制流程。

2. 材料和方法

从 2022 年 3 月至 11 月，对 32 家法国地区和大学医院进行了访谈，包括现有的和未来的 CDW。

2.1 伦理声明

这项工作已获得法国卫生高级管理局（HAS）董事会的授权。通过电子邮件要求每位受访参与者参加，并告知可能的出版形式：法国官方报告和国际出版物。此外，在每次采访中，每个参与者在录制采访之前都会被要求征得他们的同意。只有 1 名参与者拒绝录制视频。

2.2 采访

对以下主题进行了半结构化访谈:CDW 的启动和建设，项目的现状和开展的研究、机会和障碍、以及观察研究的质量标准。S1 表列出了所有受访者及其团队头衔。完整的表格和精确的问题见 S2 表。

访谈表被提前发送给参与者，然后作为进行访谈的支持。访谈时间为 90 分钟，录音供参考。

2.3 定量方法

三个表详细介绍了 S1 文本中的结构化答案。前 2 个表处理参与者和数据仓库的特征。我们根据采访期间的笔记、录音以及要求参与者提供更多信息来完成它们。第三张表侧重于 CDW 正在进行的研究。我们从专门的报告门户网站收集了这些研究的列表，我们在 8 个运营 CDW 中的 14 个中找到了这些研究。我们根据 OHDSI 研究网络描述的回顾性研究的类型学制定了研究分类。我们通过将其与收集的研究进行比较来丰富这种类型学，从而得出以下 6 个类别：

- **结果频率**：医学上明确定义的目标人群的发病率或患病率估计。
- **总体表征**：一组特定协变量的特征。可行性和预筛选研究属于这一类。
- **危险因素**：确定与明确定义的临床目标（病程、护理事件）最相关的协变量。这些研究着眼于关联研究，但没有量化因素对感兴趣结果的因果效应。
- **治疗效果**：评估明确定义的干预措施对特定结果目标的影响。这些研究旨在揭示这 2 个变量之间的因果关系。
- **•诊断和预后算法的开发**:基于给定患者的临床数据，改进或自动化诊断或预后过程。这可以采取风险、预防性评分或实施诊断辅助系统的形式。这些研究是个体化医学方法的一部分，目的是在个体患者档案的水平上推断相关信息。
- **医学信息学**：方法论或工具导向。这些研究旨在提高研究人员和临床医生的理解和行动能力。它们包括决策支持工具的评估、从非结构化数据中提取信息或自动表型方法。

研究根据其标题和描述根据该命名法进行分类。

3. 结果

图 2 总结了法国 CDW 的发展进展。在法国的 32 家地区和大学医院中，14 家正在生产 CDW，5 家正在试验中，5 家有潜在的 CDW 项目，8 家在撰写本文时没有任何 CDW 项目。所有至少处于预期阶段的项目的结果都进行了描述，减去我们在多次提醒后无法采访的 3 个项目（奥尔良、梅斯和卡昂），得出 21 家大学医院的分母。

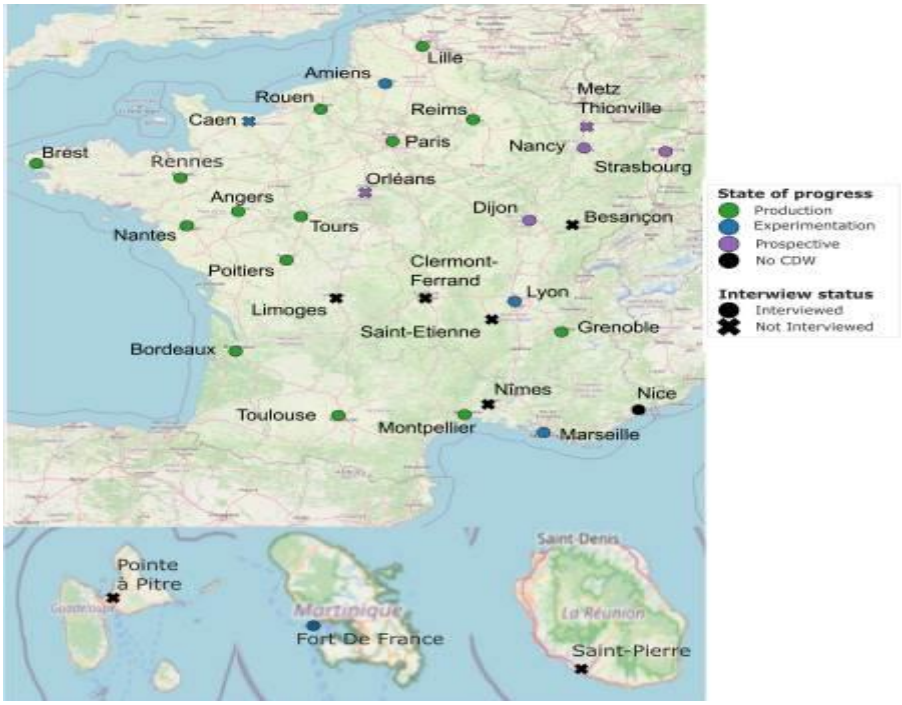

图 2 法国 CDW 的重新划分

底图和数据来自 OpenStreetMap 和 OpenStreetMap Foundation。链接到地图的基础图层：https://github.com/mapnik/mapnik。CDW，临床数据仓库。

3.1 统辖

图 3 显示了 CDW 的实施历史。必须区分第一批作品（蓝色），它系统地先于法国信息技术和自由委员会（CNIL）的监管授权（绿色）。

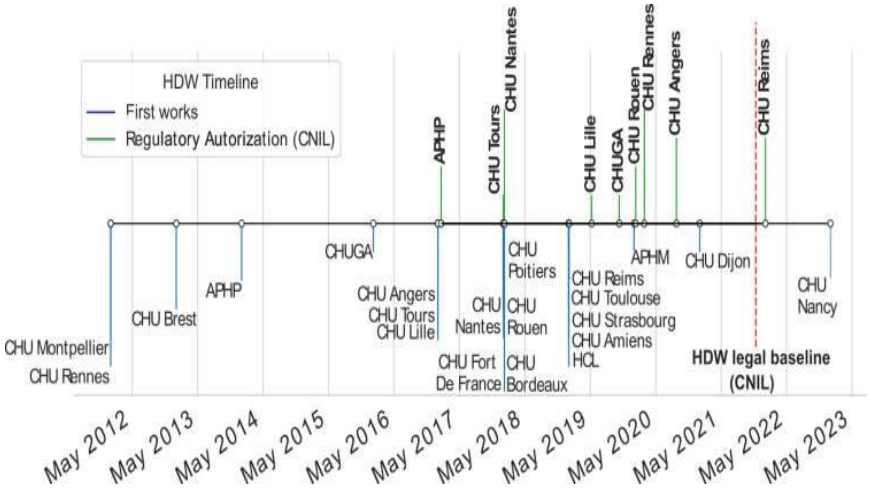

图 3 法国 CDW 的实现可以追溯到 2010 年代初数据重用的第一批学术著作，最近加速了

到目前为止，CDW 是由来自医院界的 1 或 2 名具有生物信息学、医学信息学或统计学学术背景的人发起的。CDW 的可持续性伴随着不同参与者之间合作环境的建

设：医疗信息部（MID），信息系统部（IT），临床研究部（CRD），临床用户以及管理层或机构医学委员会的支持。它还伴随着一个团队或实体的创建，致力于维护和实施 CDW。最近的举措，例如 HCL（里昂市医院）或*大东部*地区的举措，以初始，机构和高级别支持而著称。

CDW 在 CRD，IT 部门和 MID 的积极参与下，对医院的不同业务部门具有联合潜力。虽然总有一个可操作的 CDW 团队，但分配给它的人力资源差异很大：从 AP-HP 的一半全职相当于 80 人，中位数为 6.0 人。该团队系统地包括一名协调医生。它是多学科的，具有公共卫生，医学信息学，信息学（Web 服务，数据库，网络，基础设施），数据工程和统计学方面的技能。

从历史上看，第一批 CDW 基于内部解决方案开发。最近，私人行为者正在为 CDW 的实施和实施提供服务（15/21）。这些服务的范围从建立数据流和数据清理的技术专长到集成数据处理不同阶段的平台的交付。

3.2 研究管理

在开始之前，项目由科学和伦理委员会进行系统分析。经常提到本地提交和跟进平台（12/21），但其功能范围没有明确定义。它的范围从项目的简单授权到将数据自动提供给可信研究环境（TRE）。在 CDW 上启动新项目的流程总是在内部进行沟通，但很少公开记录（8/21）。

3.3 透明度

在医院网站上被公开引用的正在进行的 CDW 研究分布并不均等。一些机构有全面的研究，而另一些机构在他们的公共网站上只列出了十几项研究，而在采访中支提到了数百个正在进行的项目。总的来说，我们在产生的 14 个 CDW 中发现了 8 个这样的门户。除了正在进行的科学研究之外，很少有文献记录。正在进行的研究清单的发布是非常不一致的，并且分散在几个来源之中:临床治疗官网、健康数据中心强制性项目网或医院数据仓库网。

3.4 数据

对 HIS 有很强的依赖性。CDW 数据反映了医院工作人员每天使用的健康信息系统。利益相关者指出，CDW 数据的质量和快速有效重用所需的工作量高度依赖源 HIS。从 HIS 中以结构化和标准化格式访问数据的可能性大大简化了其与 CDW 的集成，然后简化了其重用。

数据类别

尽管全国各地的软件环境各不相同，但 HIS 的主要功能是相同的。因此，我们可以根据卫生信息系统中常见数据的主要类别，对化学武器的内容进行分析。

所有 CDW 的公共基础由来自患者行政管理软件的数据(患者身份、医院移动)和计费代码组成。然后，组成 HIS 的各种软件逐步开发数据流。目标是构建一个同构的数据模式，将数据源连接在一起，由 CDW 团队控制。资源的优先次序是通过主题项目来完成的，这些主题项目为 CDW 的建设过程提供支持。这些项目通过让 CDW 团队面对数据中存在的质量问题，提高了对相关数据源的理解。

表 1 给出了法国 CDWs 中数据类别的不同比例。结构化生物学和课本几乎总是整合在一起的(20/21 和 20/21)。这些课文包含了大量的信息。它们构成非结构化数据，因此比结构化表更难使用。其他综合来源是医院药物循环(处方和给药，16/21)，重症监护病房(ICU, 2/21)或护士表格(4/21)。成像很少集成(4/21)，主要是由于体积的原因。基因组数据被很好地识别，但从未被整合，尽管它们有时被认为是重要的，并被包括在 CDW 工作计划中。

表 1 集成到法国 CDW 中的数据类型

数据的类别	车辆碰撞车数量	率
行政	21	100%
帐单代码	20	95%
生物学	20	95%
文本	2	95%
药物	16	76%
成像	4	19%
护士表格	4	19%
解剖病理学	3	14%
重症监护室	2	10%
医疗器械	2	10%

数据复用

现 CDW 的主要用途是科学研究。这些研究主要是观察性的（非干预性的）。图 4 介绍了在 6 家医院的研究门户网站上收集的 231 项研究的定量方法中定义的 9 个类别的分布。这些研究首先关注人群特征（25%），其次是决策支持过程的发展（24%），风险因素的研究（18%）和治疗效果评估（16%）。

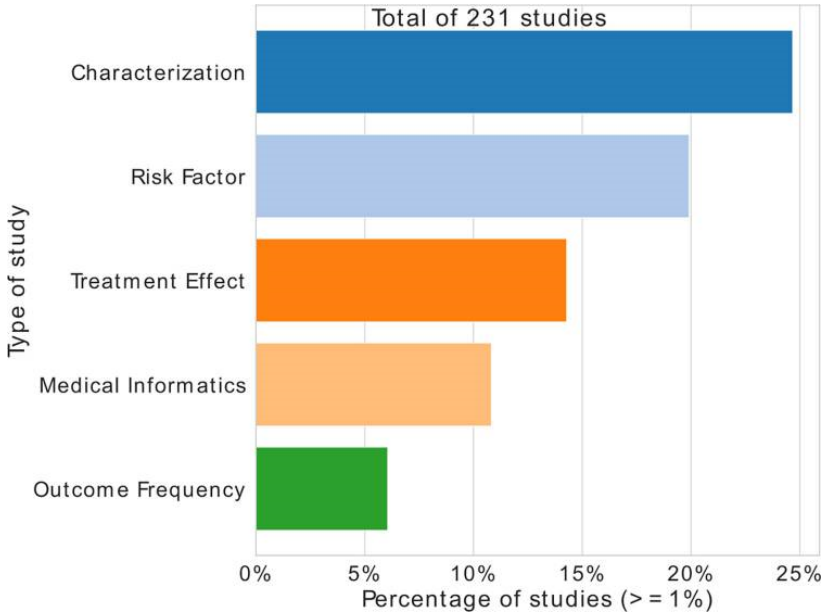

图 4 按目标划分的研究百分比

CDW 广泛用于内部项目，如学生论文（至少在 9/21），并作为单一服务研究的基础设施：它们最大的兴趣是不同信息系统的去孤岛化。对于大多数受访机构来说，仍然缺乏资源和成熟的方法和工具来进行机构间研究（例如在法国的大*西部*地区）或通过欧洲项目征集（EHDEN）。这两个研究网络分别通过超本地治理和通用数据模式 eHop 和 OMOP 成为可能。巴黎医院由于其区域覆盖和 OMOP 的选择，在多中心研究方面也非常先进。同时，*Grand-Est 地区正在基于 Grand-Ouest* 地区的模型构建 CDW 网络，也使用 eHop。

CDW 用于监控和管理 （16/21）

CDW 有时是为了改进和优化计费编码而启动的（4/21）。在同一数据库中收集的临床文本使用关键字进行查询，以方便信息的结构。然后将数据汇总成指标，其中一些指标在国家一级报告。从临床数据中构建指标也可用于机构的行政管理。最后，在接近诊所的地方，一些参与者表示，CDW 也可用于向医疗保健专业人员提供有关其实践的定期和适当的反馈。这些反馈将有助于提高医疗保健专业人员对 CDW 项目的参与度和兴趣。CDW 有时对健康监测（例如，在 COVID-19 期间）或药物警戒（13/21）感兴趣。

在护理环境中对 CDW 的浓厚兴趣（13/21）

一些 CDW 开发特定的应用程序，与护理软件相比，这些应用程序提供了新功能。搜索引擎可用于查询 CDW 中收集的所有医院数据，而无需在不同软件之间进行数据划分。然后，专用接口可以提供患者数据历史的统一视图，具有专业间的横向性，这在内科医学中特别有价值。这些跨学科检索工具还使医疗专业人员能够在所有文本中进行快速检索，例如，找到相似的患者[32]。还强调了预防、重复性任务自动化和护理协调的用途。具体的例子是按复杂程度顺序自动分类医院处方或建立初级或二级预防的专门渠道。

3.5 技术架构

现代 CDW 的技术架构有三层：

- 数据处理：源数据的连接和导出，多样化的转换（清洗、聚合、过滤、标准化）。
- 数据存储：数据库引擎、文件存储（在文件服务器或对象存储上）、用于优化某些查询的索引引擎。
- 数据暴露：原始数据、API、仪表板、开发和分析环境、特定 Web 应用程序。

补充的跨职能组件可确保平台的高效和安全运行：身份和授权管理、活动日志记录、服务器和应用程序的自动化管理。

分析环境（Jupyterhub 或 RStudio 数据实验室）是该平台的关键组件，因为它允许在 CDW 基础设施内处理数据。在我们研究时（6/21），一些 CDW 拥有这样的操作数据实验室，几乎所有 CDW 都决定将其提供给研究人员。目前，临床研究团队仍然经常在不太安全的环境中进行数据提取。

3.6 数据质量，标准格式

质量工具

一些 CDW 正在构建系统的数据质量监控流程。 通常（8/21）脚本定期运行，以检测数据流中的技术异常。内部开始开发仪表板形式的稀有数据质量调查工具

（3/21）。关于自动化数据一致性检查的可能性的理论思考正在进行中，例如人口统计或时间。一些设施从 EHR 中随机提取记录，以将其与 CDW 中的信息进行比较。

标准格式

没有一个单一的标准数据模型被所有 CDW 使用。所有人都知道 OMOP（研究标准）和 HL7 FHIR（通信标准）模型的存在。一些 CDW 认为 OMOP 模型是仓库的核心部分，特别是用于研究目的（9/21）。欧洲呼吁 EHDEN 项目鼓励了这一趋势，该项目由 OHDSI 研究联盟发起，OHDSI 是该数据模型的发起者。在法国大*西部*地区，CDW 使用 eHop 仓库软件。后者使用也称为 eHop 的通用数据模型。该模型将随着 *Grand Est* 地区未来的仓库网络也选择此解决方案而扩展。包括该组和其他选择 eHop 的机构，该模型包括 12 所大学医院中的 32 家。这使得 eHop 采用者能够启动雄心勃勃的区域间项目。但是，eHop 没有定义在其模型中使用的标准命名法，并且与新兴的国际标准不一致。

文档

一半的 CDW 已经制定了可在组织内访问的有关数据流，合格数据的含义和正确使用的文档（10/21 提到）。此文档由开发和维护仓库的团队使用。用户还使用它来了解对数据执行的转换。但是，它永远不会公开。数据在转换并准备用于分析后，不会发布任何架构。

4. 讨论

4.1 主要发现

我们首次概述了法国大学医院的 CDW，并审查了 32 家医院。CDW 的实施始于 2011 年，并在 2020 年底加速实施。目前有 24 所大学医院正在进行 CDW 项目。从本案例研究中可以得出一些通用影响因素，这些影响因素对于在全国范围内实施 CDW 的所有医疗保健系统都应该是有价值的。

4.2 管辖

随着 CDW 成为医院数据管理的重要组成部分，应鼓励建立一个致力于数据架构、过程自动化和数据文档的自主内部团队。这个多学科团队应该对数据收集过程

和潜在的重用有很好的了解，以便鉴定来自源 IS 的不同流，将它们标准化为同质模式并协调语义。它应具备充分的公共卫生知识，以及开发促进数据重用的高质量软件的技术和统计技能。

特定于仓库的资源很少，通常取自其他预算或基于项目的积分。虽然这对于初始原型制作阶段来说是很自然的，但它似乎并不适应该工具的常年和横向性质。作为一个日益重要的研究基础设施，它必须拥有长期规划的财务和组织手段。

CDW 的治理有多个层面：大学医院内部的本地，区域间和国家/国际。第一级允许确保数据集成的质量以及临床医生自己重复使用数据的相关性。区域间一级非常适合资源共同化和协作。最后，国家和国际层面确保协调，鼓励就元数据或互操作性等承诺选择达成共识，并提供财政、技术和监管支持。

4.3 透明度

卫生技术评估机构提倡在进行分析之前公开注册比较观察性研究方案。他们通常将临床治疗官网称为观察性研究的潜在但不理想的注册门户。研究界主张对所有观察性研究进行公开注册。最近，它强调需要更容易的数据访问和研究代码的发布。我们接受这些建议，并指出法国这些研究报告制度的不幸重复。一个来源可以在国家一级得到青睐，第二个来源可以通过商定共同的元数据自动从参考来源提供。

从患者的角度来看，目前无法知道他们的个人数据是否包含在特定项目中。需要更好的患者信息来重新利用他们的数据，以建立长期的信任。严格的最低限度是建立和更新每个机构正在进行的研究的声明性门户。

4.4 数据和数据应用

使用 CDW 时，分析师尚未定义数据收集过程，并且通常不知道记录信息的上下文。医学研究的这一新维度需要数据科学技能的更大发展，以将重点从统计设计的实施转变为数据工程过程。数据重用需要更多的精力来准备数据并记录执行的转换。

HIS 系统的异构性越强，在其上构建的 CDW 的定性就越低。需要提高互操作性，以帮助 EHR 供应商连接不同的医院软件，从而促进 CDW 的开发。朝这个方向迈

出的一步是开源发布 HIS 数据模式和词汇表。在分析层面，国际建议坚持需要通用的数据格式。然而，仍然缺乏采用医院 CDW 的研究标准来跨多个地点进行强有力的研究。在这些标准（如 OHDSI）之上构建开源工具可以促进它们的采用。最后，在许多临床领域，如果没有国际数据共享合作，很难获得足够的样本量。因此，需要更多的煽动来维持和更新地方术语与国际标准之间的术语映射。

许多正在进行的研究涉及决策支持流程的发展，其目标是为医疗保健专业人员节省时间。这些通常是研究项目，尚未纳入常规护理。对研究门户网站和访谈的分析表明，面向初级保健的数据重用仍然很少，很少得到适当资金的支持。从研究到临床实践的转化需要时间，需要长期支持才能产生实质性结果。

4.5 技术架构

CDW 的工具、方法和数据格式由于强大的技术创新和众多参与者的存在而缺乏协调性。正如最近关于英国使用数据进行研究的报告所建议的那样，明智的做法是关注少数模型技术平台。

这些平台应该支持开源解决方案，以确保默认的透明度，促进协作和共识，并避免医院的技术锁定。

4.6 数据质量和文档

质量本身并没有被充分考虑为一个相关的科学主题。但是，它是 CDW 内所有研究的支柱。为了提高研究用途方面的数据质量，有必要对这一主题进行持续研究。这些研究应有助于反思数据质量的方法和标准工具，例如 OHDSI 研究网络开发的方法和标准工具。

最后，需要开源发布研究代码，以确保高质量的回顾性研究。最近的数据分析研究表明，训练数据集中潜伏着无数的偏差。数据模式的开放发布被认为是所有数据科学和人工智能使用不可或缺的先决条件。受数据集卡和数据集发布指南的启发，定义一个记录主要数据流的标准 CDW 卡会很有趣。

4.7 局限性

面谈是在有限的时间内以半结构化的方式进行的。因此，一些主题被更快地涵盖，只有参与者明确提到的主题才能被记录下来。研究门户网站的不均衡存在引入了对 CDW 上进行的研究类型的记录的偏见。那些拥有透明度门户的人已经在用例中更加成熟。

为了清楚起见，我们的结果集中在大学医院的周边。我们还没有涵盖法国详尽的医疗保健领域。CDW 倡议也存在于初级保健、小型医院集团和私营公司中。

5. 结论

法国 CDW 生态系统正在形成，这得益于国家资助、专门从事健康数据的行业参与者的增加以及对欧洲健康数据空间的超国家反思，目前 CDW 生态系统正在加速发展。然而，有些方面需要特别注意，以确保 CDW 的潜力转化为患者的利益。

目前当务之急是创建和维持能够操作 CDW 并支持各种项目的多学科仓库团队。公共卫生、数据工程、数据管理、统计和 IT 能力的结合是 CDW 成功的先决条件。该团队应是数据利用问题的特权联络点，并应与现有医院部门密切合作。

建立一个多层次的协作网络是另一个优先项。本地级别对于构建数据并理解其可能的用途至关重要。区域间、国家和国际间的协调将使设立专题工作组成为可能，以便激发合作和互助的动力。

应该鼓励一种通用的数据模型，使用精确的元数据来映射集成的数据，以便限定今天从 CDW 开发的用途。更广泛地说，为提高质量而进行的数据流和转换的开源文档需要更多的激励措施，以释放所有卫生数据重用者的创新潜力。

最后，必须提出将数据范围扩大到纯医院领域之外的问题。CDW 缺少许多危险因素和患者随访数据，但对了解病理至关重要。将城市数据和医院数据结合起来，将提供一个完整的病人护理视图。