临床电子健康记录数据生命周期中数据质量分析透明报告框架

Melinda WASSELL [a,1], Kerryn BUTLER-HENDERSON [b], Karin VERSPOOR [a]
[a] 皇家墨尔本理工大学计算技术学院，澳大利亚墨尔本
[b] 查尔斯特大学护理、护理辅助与健康科学学院，澳大利亚新南威尔士州。
[1] drmelinda.wassell@gmail.com
ORCiD ID: Melinda Wassell https://orcid.org/0000-0002-7139-3164

摘要。 数据质量和二次利用数据的透明度是延缓临床人工智能模型应用、影响临床医生对其信任的关键因素。许多数据质量研究未能明确生命周期中质量检查的发生环节，导致数据溯源和再利用适用性存在不确定性。本研究制定了一个框架，用于对临床电子健康记录（EHR）数据生命周期各阶段的数据质量评估进行透明报告。该报告框架通过迭代分析制定，以确定临床数据生命周期中的参与者和阶段。框架区分了数据生成组织和数据接收组织，使用户能够将数据质量参数映射到数据生命周期的各个阶段。该框架定义了 5 个关键生命周期阶段和多个参与者。将该框架应用于真实世界数据集后，结果表明它能够揭示数据质量问题可能的来源。该框架为报告数据质量评估提供了一种结构化方法，可提高数据再利用适用性的透明度，从而支持可靠的临床研究、人工智能模型开发和内部组织治理。本研究为研究人员理解数据溯源提供了实用指导，也帮助组织在数据生命周期各阶段有针对性地开展数据质量改进工作。

关键词。 数据质量、电子健康记录、数据生命周期、透明度、数据溯源

1. 引言

随着自然语言处理技术的进步以及结构化数据收集的普及，医疗领域可供再利用的电子健康记录数据量不断增加。这一扩展带来了更多机会，使数据质量问题可能在临床电子健康记录数据生命周期的多个环节出现 (1)。每一个新的数据来源以及每一次与数据的交互，都可能成为质量问题的潜在诱因 s，进而损害数据的再利用适用性。如果不提高数据质量报告的透明度，数据增多只会带来更多质量问题。

人工智能 (AI) 在医疗领域的迅速应用受到数据质量和可信度相关障碍的制约；人们担忧，基于数据溯源或质量不明的数据构建的模型或研究是否可靠 (2)。

互操作性和标准化数据框架（例如澳大利亚互操作性核心数据 (AUCDI)(3)）正在推动健康数据共享加速发展。当患者医疗团队中的多名从业人员通过多个系统生成数据时，必须具备透明的质量评估和报告机制，以便追溯数据溯源并增强对数据的信心。

此外，医疗机构越来越多地利用自身的临床数据开展内部质量与安全改进、临床决策支持，以及

以发展学习型卫生系统（4, 5）。随着医疗保健转向数据驱动的护理，管理人员和资助机构等新的利益相关者也随之出现，他们依赖准确、及时的信息来支持患者护理和组织治理。

许多利用电子健康记录进行建模的研究仅涉及有限的数据质量参数，例如完整性。数据质量框架定义了应评估哪些内在、情境和技术参数（6），但没有说明这些评估应在生命周期的哪个阶段进行。尽管生命周期概念已得到确立（7, 8），现有框架却未能提供一种结构，以系统地分类并报告这些评估发生在数据生命周期的哪个阶段，以及由哪类参与者执行。大多数数据质量分析关注的是数据再利用时的数据，即数据经过转换以用于登记系统或通用数据模型之后的数据（9）；此时，数据工程师可能难以了解影响最初数据收集的因素。尽管已有研究指出了在数据生成阶段进行数据质量评估的重要性，但组织间的差异使这项工作颇具挑战（10）。

如果缺乏清晰的溯源信息和基于生命周期的质量报告，研究人员、人工智能开发人员和组织利益相关者就无法充分评估数据是否适用于预期用途，也无法追溯质量问题的根源。因此，透明报告整个生命周期中数据质量评估的开展位置，对于支持数据再利用和增强对二次数据使用的信任至关重要。因此，本文将介绍一种框架的开发，该框架旨在提高临床电子健康记录数据生命周期中数据质量评估的透明度。

2. 方法

我们调整并扩展了临床数据二次使用生命周期模型（8），构建了一个新的报告框架，用于报告临床电子健康记录数据生命周期中需要进行数据质量评估的环节。研究采用迭代流程，邀请临床医生和技术专家参与，以解决利益相关者冲突。我们评估了从临床电子健康记录数据最初生成到二次再利用的全过程，以记录数据创建或修改的节点。

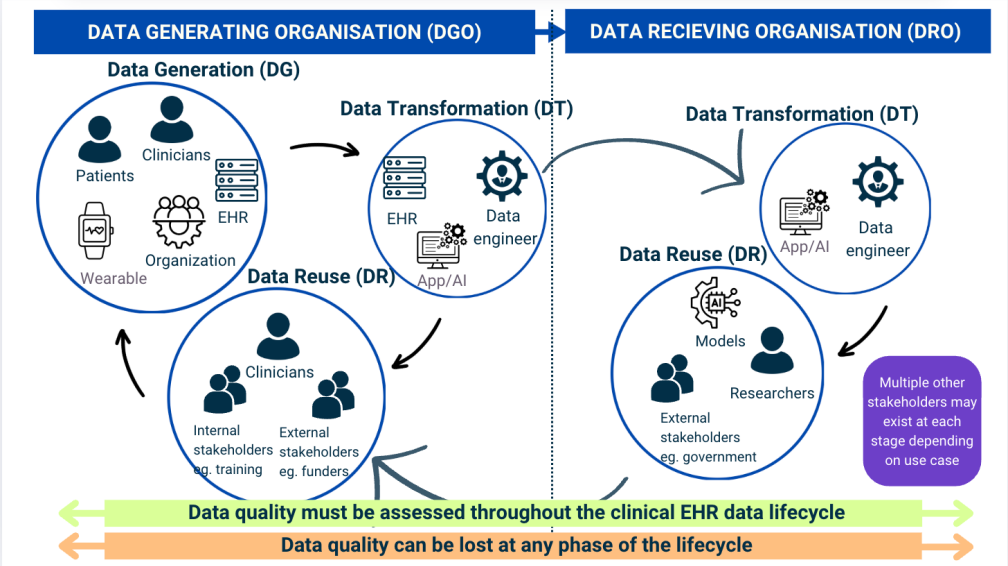
本地知识专家首先确定了两个广义的生命周期阶段：数据生成组织（DGO），即临床数据的来源组织；以及数据接收组织（DRO），通常为登记库或研究组织。随后，他们进一步细分生命周期阶段，并定义每个阶段的参与者。

为协助系统地开发并尽早应用该框架，我们选用了一个经过全面数据质量分析的电子健康记录数据集（11），通过应用该模型，确保生命周期阶段和参与者均为适合纳入的节点。该电子健康记录数据集来源于澳大利亚的职业健康诊所，包含57,570 条患者伤害记录。该组织已建立多项完善的数据使用治理流程。电子健康记录数据主要通过结构化数据采集获得，该数据集已被确认为与澳大利亚劳动力和职业伤害相关（12）。框架开发完成后，我们将潜在的数据质量检查映射至扩展数据质量评估框架中定义的参数，以确保其适用性。伦理审批：皇家墨尔本理工大学 # 26603。

3. 结果

框架开发过程最终形成了一个结构化模型，用于在临床电子健康记录数据生命周期中透明报告数据质量评估（图 1）。该框架定义了五个关键生命周期阶段，并明确了各阶段的责任参与者。

图 1. 用于透明报告数据质量分析的临床电子健康记录数据生命周期阶段。



在 DGO 中，纳入了 3 个生命周期阶段：数据生成、数据转换和数据再利用。我们的模型描述了与用于验证的数据集相关的各阶段参与者。患者、临床医生和可穿戴设备都可以直接生成数据。电子健康记录系统通常也会生成自身的数据，例如时间戳和计算结果。组织是一个独立的参与者，因为组织通常会制定管理临床医生数据录入的流程和政策，而这些可能影响数据质量或导致录入偏倚。例如，某些代码可能符合报销条件，因此临床医生可能会从几个相似的代码中选择一个，以确保其适用性。电子健康记录培训流程也可能对数据质量产生重大影响，因此在数据质量报告中将其标注为此类因素是合适的。

由于临床医生和医疗机构负责录入记录的数据，因此可穿戴设备被视为DGO-DG阶段的参与者。这一区分至关重要，因为患者使用的许多设备可能未经监管批准，因而可能成为重要的错误来源。明确标示临床读数是由临床医生测得，还是由软件应用程序自动生成，对于数据再利用者了解潜在错误至关重要。

数据转换（DT）是指将数据转换为所需格式或进行计算的流程。参与者通常是数据工程师，相关操作也可能由电子健康记录系统自动完成。

数据再利用是指数据再利用者（例如临床医生、研究人员或利益相关者）进行数据质量检查的阶段。参与数据再利用的利益相关者包括外部资助机构或政府组织。基于电子健康记录临床数据构建的人工智能模型或临床决策支持工具也属于这一阶段。随着学习型卫生系统不断发展，数据可能经历多次迭代，因此明确再利用阶段的参与者变得愈发重要。

数据转换和数据再利用都发生在DGO内部及DRO处，因为这两个位置各自存在不同的用例，而DRO往往无法了解影响数据生成的组织知识和情境因素。

根据现有框架（6，10），将数据集的数据质量检查按照内在（完整性、符合性和合理性）、情境（及时性、相关性、可访问性和治理）以及系统/技术（运行平台和互操作性）数据质量参数进行分类。

该框架显示，治理和及时性等情境参数主要在 DGO 层级进行评估，不过，不同参数在 DGO 层级和 DRO 层级的相关性可能有所不同。该框架还表明，可能需要在不同的生命周期阶段或针对不同的参与者开展类似的数据质量检查，从而揭示单次评估可能遗漏的质量问题。

4. 讨论

生命周期报告框架使再利用者能够说明数据质量检查的开展位置，以及哪些参与者参与了数据的创建和修改。该框架通过提供描述数据历程的标准化语言，支持提高数据溯源的透明度，并建立对人工智能模型的信任。基于生命周期的报告有助于识别更具体的数据质量问题领域，并指出干预措施应针对临床医生培训、组织政策还是系统设计。

4.1. 与现有数据质量框架的关系

该框架补充了现有的数据质量框架和工具。扩展并协调统一的内在数据质量框架 (6, 10) 定义了要评估哪些参数，而我们的框架则明确了评估发生在哪些环节。Achilles (13) 等工具可提供自动化评估，而我们的框架则为结果报告提供语义背景。例如，若将 Achilles 报告中“15% 的诊断代码缺失”这一结果表述为：“DGO- DG-Organization（政策：规定仅在可计费就诊时才需要诊断）”以及“DRO-DT-Engineer（映射：成功率为 92%）”，便能揭示政策和技术因素对缺失情况的影响。 l

该框架符合数据溯源文档方面的新兴要求 (14)，例如 FAIR（可查找、可访问、可互操作、可再利用）原则 (15)。生命周期术语可纳入元数据文档，以支持机器可读的溯源跟踪。此外，在内部发生再利用时，识别组织参与者（DGO-Org、DGO-Clinician）有助于内部治理和问责。

4.2. 实际实施与建议的报告方式

目前尚无通用标准可用于确定数据是否适用，因此数据是否恰当取决于用户及其所处的情境（16）。这可能导致不同评估者对数据是否适用得出显著不同的结论。尽管这一生命周期框架无法解决这一挑战，但它为系统、透明地判定数据是否适用提供了必要的基础设施，使数据再利用者能够评估所执行的质量检查是否符合其特定的用例要求。

该框架支持采用标准化记法进行数据质量报告。这种记法有助于提高透明度，例如：“DGO-DG-Clinician（完整性：94%）”与“DRO-DR-Researcher（完整性：87%）”相区别。这有助于明确，7%的差异反映的是研究人员采用的标准，或数据转换中的问题，而非临床医生记录方面的缺漏。例如，对于共享护理团队中的临床医生，及时性可能要求数据能够即时获取，以确保数据适用；而研究人员可能只需每月更新数据，即可判断数据是否适用。

将该框架应用于案例研究数据集中的数据质量检查分类后发现，同一项数据质量检查可能需要在不同的生命周期阶段或针对不同的参与者开展。例如，时间合理性是指数据是否可信。在组织层面进行评估时，可能存在一项政策，要求所有临床医生都必须报告损伤的身体侧别；因此，这些身体侧别数据适合重用。相反，在 DG-Clinician 层面进行检查时，会发现有些临床医生从不记录身体侧别，或总是记录同一侧，因此这些数据不适合重用。这表明，该框架能够揭示较粗略的评估可能遗漏的质量问题。

针对模糊案例，必须为每种数据再利用情形制定决策规则。我们建议，为确保透明度，报告应基于对数据质量影响最大的因素被认为源自何处

。例如，如果电子健康记录系统强制使用某个代码或基于列表的术语，那么出现的问题属于 DG-EHR 问题。但如果临床医生可以自由输入任何代码，质量则由 DG-Clinician 负责。实施该框架的组织应制定基于共识的指南，以适应各自的电子健康记录系统和工作流程，并记录这些决策规则，以确保在各项目中得到一致应用。

5. 局限性与未来研究

该框架以单个数据集作为案例研究开发而成，仍需进一步完善，以适用于不同的医疗保健环境。生命周期各阶段的数据质量检查分类有时需要进行主观判断，这一点应在不同组织中进一步探讨。该框架为报告检查发生的位置提供了结构，但并未规定应执行哪些检查，因为这些决定仍取决于具体情境。

该框架并不旨在描述每个环节中的所有参与者。随着医疗卫生互操作性的不断发展，患者数据收集将涉及更多参与者，例如照护者。该框架可以相应扩展。

6. 结论

该框架通过系统识别生命周期各阶段的参与者，提升了数据质量评估的透明度，并明确了数据溯源。它弥合了全面的数据质量评估框架与实际实施之间的差距，可供数据生成组织内部及外部的数据再利用者使用。随着健康数据共享在全球范围内不断扩大，该框架为透明报告数据质量特征提供了结构，支持提升研究质量和可靠的人工智能模型开发，并通过内部治理增强患者安全、临床医生质量和利益相关方报告。该框架为数据再利用者更好地理解数据的再利用适用性提供了实用指导，也帮助组织在临床数据生命周期内开展有针对性的数据质量改进工作。

致谢

MW 感谢澳大利亚政府研究培训计划提供的学费资助，并感谢 Work Healthy Australia 提供研究数据集

参考文献

1. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, 等. 电子健康记录数据质量评估及工具：系统综述。J Am Med Inform Assoc. 2023;30(10):1730-40。2. Hassan M, Kushniruk A, Borycki E. 医疗保健中的人工智能应用：障碍与促进因素——范围综述。JMIR 人因. 2024;11:e48633。3. Sparked AU FHIR 加速器。《澳大利亚临床数据互操作性》第2版。赫斯顿，昆士兰州；2025年。4. Enticott J, Johnson A, Teede H. 利用数据推动医疗保健改进和影响的学习型卫生系统：系统综述。BMC 卫生服务研究. 2021;21(1):200。

5. Foley TJ, Vale L. 学习型卫生系统在医疗服务提供者质量改进中发挥什么作用? Learn Health Syst. 2017;1(4):e10025。6. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, 等. 电子健康记录数据二次利用的协调化数据质量评估术语与框架。EGEMS (Wash DC). 2016;4(1):1244。7. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. 初级保健电子健康记录数据使用与再利用中可能存在的偏倚来源。J Med Internet Res. 2018;20(5):e185。8. Weng C. 临床数据质量：数据生命周期视角。生物统计学与流行病学（英国牛津）。2020;4(1):6-14。9. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. 通过评估观察性数据质量，提高对真实世界证据的信任。美国医学信息学会杂志：JAMIA. 2021;28(10):2251-7。10. Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, 等. 数据生命周期中真实世界数据仓库的质量评估：文献综述。J Am Med Inform Assoc. 2021;28(7):1591-9。11. Wassell M, Butler-Henderson K, Verspoor K. 电子健康记录数据再利用的数据质量报告（审稿中）。BMJ Health Care Inform. 2025年。12. Wassell M, Vitiello A, Butler-Henderson K, Verspoor K, Pollard H. 用于建模工作相关肌肉骨骼疾病结局的肌肉骨骼治疗师电子健康记录的可推广性。J Occup Rehabil. 2025;35(1):125-38。13. Huser V, DeFalco FJ, Schuemie M, Ryan PB, Shang N, Velez M, 等. 患者级临床数据集数据质量工具的多中心评估。EGEMS (Wash DC). 2016;4(1):1239。14. Johnson KE, Kamineni A, Fuller S, Olmstead D, Wernli KJ. 电子健康记录数据的溯源为何对研究至关重要：一个使用系统映射的案例。EGEMS (Wash DC). 2014;2(1):1058。15. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, 等. FAIRsharing：面向标准、数据仓库和政策的社区协作方法。Nat Biotechnol. 2019;37(4):358-67。16. Weiskopf NG, Weng C. 电子健康记录数据质量评估的方法与维度：促进临床研究中的再利用。J Am Med Inform Assoc. 2013;20(1):144-51。