Applied Medical Informatics
Review Vol. 46, Suppl. 2 /2024, pp: S41-S44

Challenges and Opportunities of Clinical Data Warehousing

Niklas GIESA[1,*], Anne Rike FLINT[1], Sedir MOHAMMED[1], Louis AGHA-MIR-SALIM[1], Sebastian Daniel BOIE[1], Fabian PRASSER[2], and Felix BALZER[1]

[1] Institute of Medical Informatics at Charité – Universitätsmedizin Berlin, Invalidenstr. 90, 10115 Berlin, Germany
[2] Medical Informatics, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany
Emails: niklas.giesa@charite.de; anne-rike.flint@charite.de; sedir.mohammed@charite.de; louis.agha-mir-salim@charite.de; sebastian-daniel.boie@charite.de; fabian.prasser@charite.de; felix.balzer@charite.de

* Author to whom correspondence should be addressed

Received: 27 June 2024 /Accepted: 11 September 2024 / Published online: 21 November 2024

Abstract
Clinical data warehousing (CDW) aims to provide an integrated database for secondary health analysis. We conducted a rapid review including 22 studies published between 2018 and 2022 that reported CWD implementation details in addition to individual experiences. Our results comprise technical details and reveal current opportunities and challenges of CDW. Many studies positively highlight standardized tools and models building on well-integrated clinical concepts. Enhanced tooling and data modeling were oftentimes hindered by bad data quality and organizational burdens. We call for enforcing synergies between clinical, technical, and organizational expertise to successfully roll out a CDW project.

Keywords: Clinical Data Warehousing; Electronic Health Records; Clinical Data Model; Rapid Review

Introduction

Clinical information systems (CIS) store high volumes of electronic health records (EHRs) [1]. To build a valuable integrated single point of data for secondary analysis, clinical data warehousing (CDW) facilitates extract, transform, and load (ETL) processes, harmonizing data from different source systems and file formats [2]. Traditional open-source projects, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) or i2b2/tranSMART provide analytical tools in addition to database schemas [3]. OMOP CDM follows a normalized design principal suggested by Bill Imnon for analytical data stores while i2b2 is structured in de-normalized star-schemas inspired by Ralph Kimball [4].

Emerging big data infrastructures, like the Hadoop [5, 6], and advanced terminology systems, like SNOMED Clinical Terms [8], have been changing the CWD landscape over the past years. Besides relational data base management systems (RDBMS) document-oriented storages like NoSQL databases have gained popularity for exchanging data in interoperable formats like Fast Healthcare Interoperability Resources (FHIR) [6, 7].

We conducted a rapid review investigating technical implementations while highlighting resulting opportunities and challenges of CDW projects. Previous reviews have focused on specific clinical domains rather than analyzing these aspects of implementation projects [2, 8]. This study aimed to provide a general conspectus guiding an early conceptual phase before the concrete CWD rollout.

Methods

We searched PubMed for publications dealing with CDW implementations. Eligibility criteria for title and abstract were CDW related search terms, such as "clinical" in the combination of "data warehouse", "data repository", "data inventory", etc.[1], yielding a total of 136 publications. Exclusions were made for references with clinical trial search terms in their title, like "electronic data capture", "case report form", "clinical trial", etc.[2], reducing the number of references to 120. After applying inclusion- and exclusion criteria for published work between 2018 and 2022, 55 references remained for full-text screening. While screening we focused on technical reports that revealed details on the used CDM and the DBMS. Subsequently, we covered 22 publications in our rapid review that we conducted instead of a systematic review to reduce time and complexity while summarizing the most important results. We screened references during 2023 and see our rapid review as an initial exploration of the complex CDW research area.

Results

Table 1 displays applied review categories and the number of references (absolute frequencies) that fall into corresponding category items, e.g., Year Published. The underlying data table including all citations is provided in our GitHub repository [9]. 31% of the included studies (7 out of 22) were published in 2019 while 2021 and 2022 make up 18% (4 out of 22) each. The number of patients covered by the CDM ranged a lot between 100 and 70 million (M) while projects tend to either include little or vast number of records.

Table 1. Review categories for 22 references denoted as [absolute frequency] × [category item]. T: thousand, M: million, RDBMS: Relational Database Management System, OMOP: Observational Medical Outcomes Partnership

Year Published	Patient Volume	Clinical Data Model	Database Technology
5 × 2018	5 × [1M-70M]	6 × i2b2	16 × RDBMS
7 × 2019	4 × [100T-1M]	4 × OMOP	3 × RDBMS + NoSQL
2 × 2020	3 × [10T-100T]	2 × Dr. Warehouse	1 × Hadoop
4 × 2021	5 × [1T-10T]	1 × i2b2 + OMOP	1 × NoSQL
4 × 2022	5 × [100-1T]	8 × non-standard	1 × Hadoop + NoSQL

The most popular standard CDM design was i2b2 (6 out of 22) followed by OMOP (4 out of 22). Rinner et al. [3] combined these two approaches integrating standardized clinical vocabulary available on the data platform Athena. Garcelon et al. [10] introduced the project Dr. Warehouse providing ETL pipelines for the population of self-designed star-schemas. Plenty of CDW projects (8 out of 22) followed a non-standard CDM, five of them stored EHRs in de-normalized formats.

73% (16 out of 22) studies implemented a RDBMS that contributes to the seamless installation of i2b2 or OMOP. In contrast, Afshar et al. [5] and Artemova et al. [6] used novel infrastructure via implementing Hadoop solutions combining structured EHRs with unstructured data like images or texts. Only one project focused on document-based NoSQL data storages [11].

Authors positively highlighted vast data integration and harmonization capabilities of different modalities overcoming disjointed data silos [5, 6, 11, 12]. Curry et al. [11] presented the combination of traffic data and EHRs. The standardized vocabulary of OMOP was seen as a valuable base for the semantic interoperable data exchange between i2b2, FHIR, and additional export formats [3, 7, 12]. Projects also enhanced existing ETL pipelines by Notebooks for R or Python enabling data quality analysis and data exploration capabilities [6, 12]. Cossin et. al. built upon the i2b2 toolset developing a front-end for EHR annotations [13].

[1]Additional inclusion search terms: ("clinical") ("data") "store", "archive", "schema", "knowledge base", "lake", "base", "platform"
[2]Additional exclusion search terms: "clinical study", "digital documentation", "electronic scan", "electronic document", "randomized control study"

Data governance and data access patterns were identified as challenging when operating a CDW [12, 14, 15]. Walters et al. comprehensively described how they dealt with data requests in compliance with data privacy regulations [14]. Fleuren et al. explained their complex pseudonymization framework [16]. Although vendors provide powerful data integration tools, semantic integration was hampered by insufficient data quality [11, 12, 15, 16]. Record linkage, as the assignment of unique identifiers to each EHR across source systems, required a lot of data exploration [15, 16]. Complex toolsets were implemented at the expense of costly end-user training [13, 14].

Discussion

We found the usage of open-source CDMs as a major opportunity of CDW implementation projects. Although i2b2 seems to be favored by developers, semantic integration is especially enforced by the OMOP CDM [3]. De-normalized design principles, advocated by Ralph Kimball [4] in the field of data warehousing research, appeared to be suitable for the central integrated storage of EHRs. Surprisingly, most of the CDW were implemented on the basis of RDBMS, a few authors describe integration aspects by novel data systems like NoSQL or Hadoop. The document-based format FHIR [7] was seen as an export format instead of a suitable CDW structure. Data governance and privacy require synchronized organizational processes with data engineering tasks [14]. We also found that semantic interoperability cannot be solved with a standard data format but demands a certain level of data quality.

The results of this work are limited to a small number of references that were available between 2018 and 2022. We aimed to provide a first basis for a more complex literature analysis. This study would have benefited from a more comprehensive systematic review sharpening our claims and arguments. In the future, we plan to analyze current CDW studies with respect to interoperability aspects of i2b2 and OMOP in the combination with FHIR.

Conclusion

Results drawn from our rapid review call for deep synergies between technical, clinical, and organizational expertise addressing data quality and governance issues while profiting from advanced open-source CDW tools.

List of Abbreviations: EHR: Electronic Health Record, CDW: clinical data warehouse, CDM: Clinical Data Model, FHIR: Fast Healthcare Interoperability Resources, RDBMS: Relational Data Base Management Systems, OMOP: Observational Medical Outcomes Partnership, HDFS: Hadoop Distributed File System, DBMS: Data Base Management System, T: thousand, M: million

Author Contributions: NG, AF, SM extracted items, screened articles. NG wrote the manuscript, LA, SB, FB proofread.

Funding: This research received no funding.

Ethics Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: None.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, Dugas M, Dupont D, Schmidt A, Singleton P, De Moor G, Kalra D. Electronic health records: new opportunities for clinical research. J Intern Med. 2013;274(6):547-60. doi: 10.1111/joim.12119.

2. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions. JMIR Form Res. 2020 Aug 27;4(8):e17687. doi: 10.2196/17687.
3. Rinner C, Gezgin D, Wendl C, Gall W. A Clinical Data Warehouse Based on OMOP and i2b2 for Austrian Health Claims Data. Stud Health Technol Inform. 2018;248:94-99.
4. Yessad L, Labiod A. Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault. 2016 International Conference on System Reliability and Science (ICSRS), Paris, France, 2016, pp. 95-99, doi: 10.1109/ICSRS.2016.7815845.
5. Afshar M, Dligach D, Sharma B, Cai X, Boyda J, Birch S, et al. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. J Am Med Inform Assoc. 2019;26(11):1364-1369. doi: 10.1093/jamia/ocz068.
6. Artemova S, Madiot PE, Caporossi A; PREDIMED group; Mossuz P, Moreau-Gaudry A. PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital. Stud Health Technol Inform. 2019 Aug 21;264:1421-1422. doi: 10.3233/SHTI190464.
7. Solbrig HR, Hong N, Murphy SN, Jiang G. Automated Population of an i2b2 Clinical Data Warehouse using FHIR. AMIA Annu Symp Proc. 2018 Dec 5;2018:979-988.
8. Hamoud A, Hashim AS, Awadh WA. Clinical data warehouse: a review. Iraqi Journal for Computers and Informatics 2018;44(2):16-26. doi: 10.25195/2017/4424.
9. Niklas G. Challanges and Opportunities of Clinical Data Warehousing. Available from: https://github.com/ngiesa/cdw_review (accessed 20/06/2024)
10. Garcelon N, Neuraz A, Salomon R, Faour H, Benoit V, Delapalme A, et al. A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. J Biomed Inform. 2018;80:52-63. doi: 10.1016/j.jbi.2018.02.019.
11. Curry AE, Pfeiffer MR, Metzger KB, Carey ME, Cook LJ. Development of the integrated New Jersey Safety and Health Outcomes (NJ-SHO) data warehouse: catalysing advancements in injury prevention research. Inj Prev. 2021;27(5):472-478. doi: 10.1136/injuryprev-2020-044101.
12. Fruchart M, Quindroit P, Patel H, Beuscart JB, Calafiore M, Lamer A. Implementation of a Data Warehouse in Primary Care: First Analyses with Elderly Patients. Stud Health Technol Inform. 2022;294:505-509. doi: 10.3233/SHTI220510.
13. Cossin S, Lebrun L, Aymeric N, Mougin F, Lambert M, Diallo G, et al. SmartCRF: A Prototype to Visualize, Search and Annotate an Electronic Health Record from an i2b2 Clinical Data Warehouse. Stud Health Technol Inform. 2019;264:1445-1446. doi: 10.3233/SHTI190476.
14. Walters KM, Jojic A, Pfaff ER, Rape M, Spencer DC, Shaheen NJ, et al. Supporting research, protecting data: one institution's approach to clinical data warehouse governance. J Am Med Inform Assoc. 2022;29(4):707-712. doi: 10.1093/jamia/ocab259.
15. Visweswaran S, McLay B, Cappella N, Morris M, Milnes JT, Reis SE, et al. An atomic approach to the design and implementation of a research data warehouse. J Am Med Inform Assoc. 2022;29(4):601-608. doi: 10.1093/jamia/ocab204.
16. Fleuren LM, Dam TA, Tonutti M, de Bruin DP, Lalisang RCA, Gommers D, et al. The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. Crit Care. 2021;25(1):304. doi: 10.1186/s13054-021-03733-z.