Journal of the Society for Clinical Data Management

Wang Z, et al. Clinical Data Warehousing: A Scoping Review. *Journal of the Society for Clinical Data Management*. 2024; 4(1): 8, pp. 1–19. DOI: https://doi.org/10.47912/jscdm.320

REVIEW ARTICLE

Clinical Data Warehousing: A Scoping Review

Zhan Wang*, Mahanaz Syed*, Shorabuddin Syed*, Melody Greer†, Emel Seker†, Meredith N. Zozus* and Catherine K. Craven*

Introduction: A clinical data warehouse (CDW) is a powerful resource that supports clinical decision-making and secondary data use by integrating and presenting heterogeneous data sources. Despite considerable effort within healthcare organizations (HCOs) to develop CDWs, scientific literature surrounding clinical data warehousing methods is limited.
Objectives: The scoping review aims to characterize the current state of CDW methods within HCOs, to identify extant evidence for practice recommendations, and ultimately to advance the design, implementation, and use of CDWs.
Methods: The review encompasses CDW articles published from 2011 through 2021 identified through a systematic PubMed search. Article abstracts were systematically screened by two authors. Full-text articles were reviewed and abstracted independently by two authors with discrepancies resolved through consensus.
Results: 137 articles, from 55 journals and 3 conference proceedings, were categorized and analyzed. Areas for increased CDW focus include CDW design (such as data integration of increased data types and sources; extract-transform-load (ETL) optimization; data quality improvement processes; semantic data representation) and CDW governance (such as support tools/documentation and data literacy efforts for staff and end-users; governance structure; and business model/financial support for CDWs including staffing).
Conclusion: The study indicates the topics that have been significantly developed and the aspects that need additional focus and reporting in CDW between existing general data management best practices and recently articulated requirements for research data. Also, more multi-site and multi-aspect studies are needed to foster maturity at CDWs.

Keywords: Data Warehousing[Mesh]; Clinical data, Data Accuracy[Mesh]; Scoping review; Informatics[Mesh]

Introduction

Electronic health record systems (EHRs) are widely used by physicians and other members of care teams in decision making. Nearly ubiquitous EHR adoption in the U.S. and elsewhere over the last decade spurred the current emphasis on the secondary use of healthcare data for research [1], and the widespread development of institutional patient data repositories for research at academic health centers and larger health systems [2, 3] A data warehouse is a collection of data that is subject-oriented, integrated, time-variant, and non-volatile that can be used to produce useful information for management decision making [4]. Organizations have developed clinical data warehouses (CDWs) to integrate, manage, and centrally provide access to EHR data (often incorporating other types of patient data) to researchers and other stakeholders.

Learning health systems require data from healthcare organizations, including internal and external clinical, operational, and financial data, which implies that healthcare data warehousing is becomming an expected capability of healthcare organizations (HCOs). In Academic Medical Centers (AMCs) where these data are also needed to support research, we would expect heightened emphasis on CDW methods. Despite considerable effort within HCOs to develop CDWs, data warehousing is not a common topic within biomedical literature. We conducted a scoping review to gain insight into the state of CDW methods, technological developments, current foci, impacts, and potential gaps from the literatures as implemented as a CDW. A scoping review, as defined by the Canadian Institutes of Health Research, is an "exploratory project that systematically maps the literature available on a topic, identifying key concepts, theories, sources of

* University of Texas Health Science Center at San Antonio, San Antonio, TX, US

† University of Arkansas for Medical Sciences, Little Rock, AR, US

Corresponding author: Zhan Wang, PhD (wangzhan0306@gmail.com)

evidence and gaps in the research [5–8]." Scoping reviews are undertaken prior to a full synthesis when a domain's literature is large, or conversely, as was the case here, when there is a lack of literature, to specify what is known [7]. We chose to search solely PubMed, the world's largest freely accessible online biomedical citation database [9], precisely because it is where the most healthcare-related journals are indexed, including those for clinical research informatics, a domain in which CDWs are a focus. The scholarly literature on CDWs at HCOs indexed within PubMed is likely to be more comprehensive than elsewhere, day [10].

The focus of the study was to achieve the following goals:

[1] Understand the current state of CDWs used in clinical organizations;
[2] Indicate CDW current improvement foci;
[3] Indicate gaps in the CDW literature between existing general data management best practices and recently articulated requirements for research data.

Material and methods

Data collection

Based on expert knowledge from the author team (CKC, MG, MNZ) as well as additional professional medical librarian input, and multiple iterative tests, we formulated the following query to represent the concept of CDW for the PubMed search:

((("Medical Records"[Mesh:noexp] OR "Medical Record Linkage"[Mesh] OR "Medical Records Systems, Computerized"[Mesh] OR "Health Information Management"[Mesh] OR "Health Record" OR "Health Record System")) AND ("Datasets as Topic"[Mesh] OR "Databases, Factual"[Mesh:noexp] OR "Database Management Systems"[Mesh] OR "Decision Making, Computer-Assisted"[Mesh] OR "Information Storage and Retrieval"[Mesh:noexp]))
OR
("data warehouse"[tiab] OR "data warehouses"[tiab] OR "data warehousing"[tiab] OR "data repository" OR "data repositories" OR "data warehousing"[Mesh] OR I2B2[tiab] OR "informatics for integrating biology and the bedside i2b2"[tiab] OR BTRIS[tiab] OR "biomedical translational research information system"[tiab] OR "star schema"[tiab]))
NOT
("animals"[MeSH Terms])

The query was designed to identify all potential literatures as implemented as CDWs, and their ongoing development, technologies employed, methodologies, and areas of study about them, over time, as language surrounding the concept matured, including when the Medical Subject Heading "Data Warehousing" was added in 2018. The query emphasized CDW methodologies and development. The articles only mentioned the use of CDW data or CDW applications were excluded. The other emphasis of the query was academic-developed or public CDWs as a result of the inclusion of i2b2 (Integrating Biology & the Bedside) terms. These emphases could lead to the under-representation of: 1) other common data model (CDM) studies (eg, Medical Outcomes Partnership (OMOP)), 2) commercial CDW studies, (eg, Epic systems, such as Clarity, Caboodle, COSMOS [11]; and other systems, such as Healtheintent, HealthFacts, etc.) 3) specific methodology studies (eg, data modeling, data quality), and 4) specific data registry studies (eg, cancer, molecular biological data, or COVID studies, such as the National COVID Cohort Collaborative). This scoping review included all relevant studies published in peer-reviewed journals indexed in MEDLINE (the largest subset of PubMed) between January 1, 2011, and September 30, 2021, when we ran our final query. We limited the years included in the scoping review only after extensive examination of the retrieved literature prior to this period, for which we deemed the content too far removed from the present to be actionable or informative outside of historical interest. The time period for the search also corresponded with the recent emphasis from the National Institutes of Health Clinical and Translational Science Award (CTSA) program – from which the first round of funding started in 2006 – on data warehousing as important clinical research infrastructure, the multi-year time period to reach large numbers of funded institutions and the lengthy design and development process for CDWs.

All citations to potentially relevant studies retrieved were also screened for inclusion [12].

Data collection and analysis

Each title, abstract, and full-text article retrieved was reviewed by two independent reviewers (CKC, ZW). When the two reviewers reached different conclusions, a third reviewer (MS or SS) adjudicated in a group discussion to produce a final decision on inclusion.

The following exclusion criteria were applied to article screening and full-text review: (1) articles from non-peer reviewed trade journals were excluded; (2) articles where the only CDW aspect reported was the using of CDW data were excluded; (3) non-English publications were excluded; and (4) articles for which the full-text was not available, abstracts and posters were excluded.

The articles included after full-text review were reviewed again (CKC, MS, SS, ZW) to abstract and categorize their content [5, 7, 12, 13]. The following information was abstracted from the included articles: article type; CDW focus; CDW design (including CDW architecture, data model, data domains included in the CDW and semantic data representation); work/improvement foci; CDW governance (including governance structure, user support tools and documentation, staff training and financial sustainability) (**Table 1**). Though started by topics that the authors, who are all data warehousing subject matter experts, knew to be emerging or of interest in the CDW community, categories and subcategories of the information abstracted from the included articles were developed iteratively throughout the review. Previously

abstracted articles were re-reviewed after new categories were added so that all information that appeared relevant and mentioned in multiple articles was abstracted. Each article was abstracted by two assigned reviewers. Reviewers recorded their abstraction work in a spreadsheet so that the presence of relevant information was systematically assessed for each article. The group of four reviewers then discussed and came to consensus decisions on articles with initial categorization differences. Iteratively, as well, through these reflexive discussions, the group discussed the varying language used across the articles and their understanding of technical topics and current foci within the CDW field, refined their understanding and definitions of categories/subcategories, and where necessary, re-abstracted for thorough, uniform categorizations until no further changes were necessary and all categorization was completed [5, 7, 12, 13].

Results

The PubMed query returned 1,763 citations for the decade spanning 2011–2021 (**Figure 1**). Ninety-eight citations were excluded in the first-round review (30 were non-English, 66 had no full-text). Another 1,527 citations were excluded in the second-round abstract review (1,092 articles were not related to CDW; 435 described only the use of data from a CDW. For the remaining 140 citations, full-text articles were reviewed by 4 reviewers, and 3 articles were excluded. Finally, 137 articles were included, abstracted, and analyzed (Appendix 1).

As shown in **Figure 2**, article numbers increased from 2011 to 2015. They dropped in 2016 after the former high in 2015, but rose again through 2021. For 2021, we collected articles until September. More articles are expected in the rest of the year during the COVID pandemic. Between 2011 and 2021, CDW-focused articles

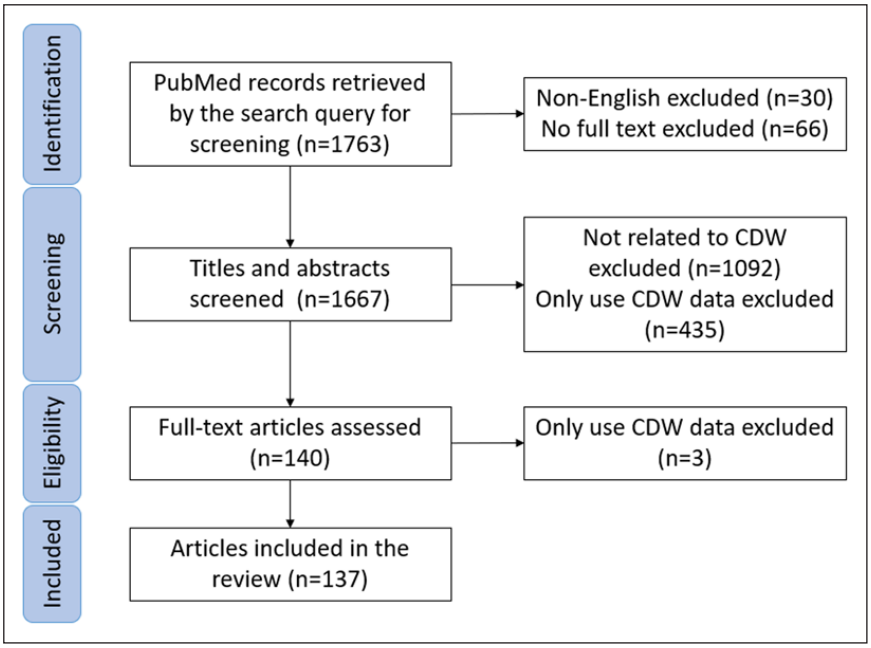

Figure 1: Flow diagram of the PRISMA Statement-based article identification, exclusion, and inclusion process.

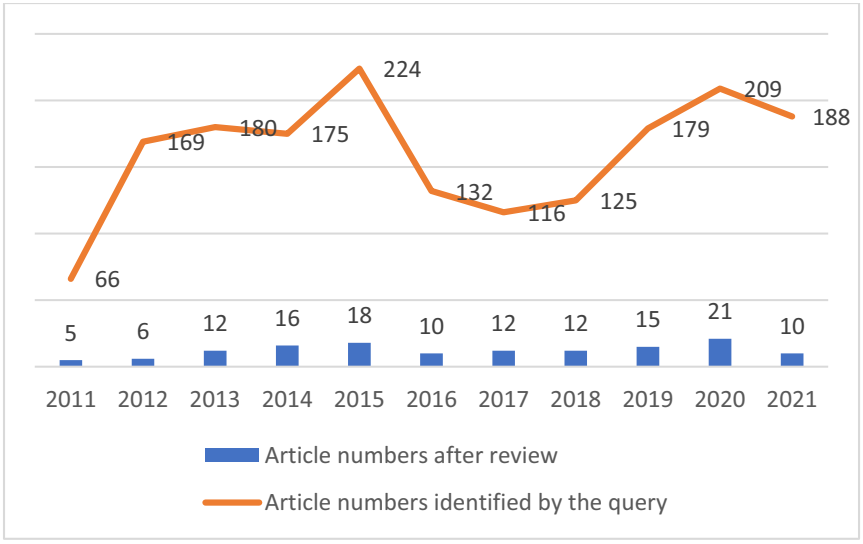

Figure 2: Ten-year trend for the numbers of CDW articles indexed in PubMed (by the query and after review).

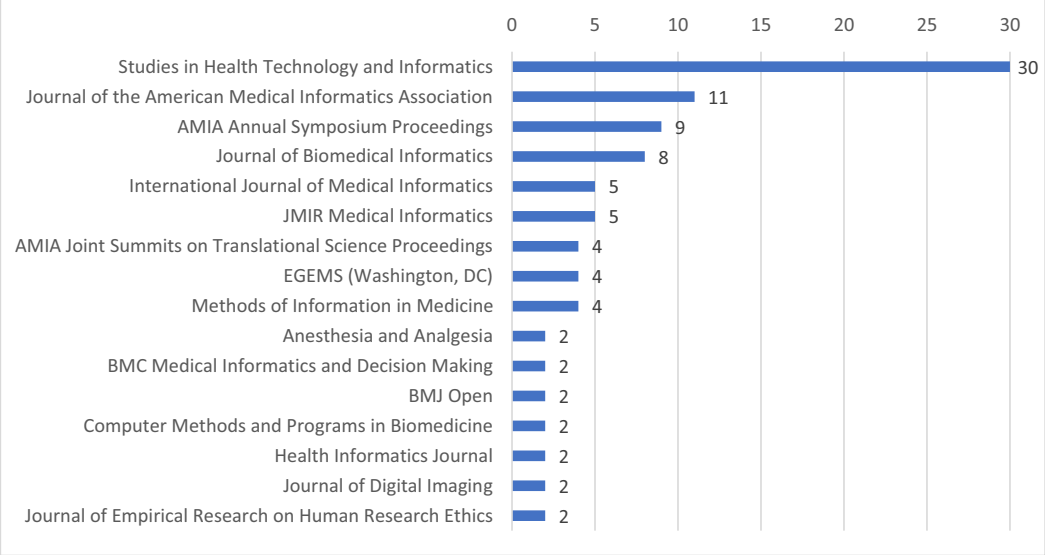

Figure 3: Publications with the most CDW articles (2011–2021).

Table 1: Themes and descriptions.

Theme	Sub-theme	Descriptions
Article type		This category describes study methodology types, including case study, qualitative study, systematic literature review, survey, and discussion. Articles that considered CDW issues but were not studies were grouped into the discussion category.
CDW focus		This category describes whether the CDW described, built, or enhanced was designed for clinical operations and decision-making, clinical research, or both.
CDW design	CDW architecture	This category describes whether the CDW's architecture was from a commercial vendor or a public architecture. Commercial architecture is used by the CDW project in partnership with industry, often from the institution's EHR vendor, or another proprietary architecture. Public architecture is developed by the CDW project at a healthcare institution or governmental agency.
	Data model	This category describes which data model was used for the CDW discussed and includes these sub-categories: CDMs, such as the Patient Centered Outcomes Research network (PCORnet), Observational Medical Outcomes Partnership (OMOP), widely used in secondary use of health care data to promote interoperability and data sharing; open-source models, available for developers to download, modify, and re-use (eg, OpenEHR); and custom models, created by a specific institution or project and not generally available for use by others.
	Data domain	This category describes the data domains presented in the articles as follows: longitudinal clinical (eg, demographic, encounter, diagnosis, vital sign, etc.); disease specific longitudinal clinical (eg, Alzheimer, obesity, neurological disease, etc.); unstructured clinical; administrative; biological molecular; imaging; and blood bank.
	Semantic data representation	This category describes ontologies or controlled terminologies used in the CDW, including existing, customized, and mapping ones.
Work/ Improvement foci		This category describes what technologies or methodologies were a primary focus of development or improvement of the CDW, based on each article's Methods section. The methodologies were comprised of the following: data integration, data warehouse implementation, data querying, data quality, use of data, data visualization, clinical decision making, standardization for sharing, data acquisition, natural language processing (NLP), protected health information (PHI).
CDW governance	Governance structure	This category describes the content that related to data governance, including data governance team build, policies, and requirements.
	End-user support and documentation	This category describes the actions for end-user support, including system user interface, data documentation, data report generation and user support team.
	Staff training	This category describes the actions for CDW staff training, including training aim, method, and plan.
	Financial sustainability	This category describes how to address costs and sustainability or raise revenue to support future work.

in the biomedical literature averaged 12.5 per year ranging from 5 to 21.

The 137 articles abstracted were published in 55 journals and proceedings from three conferences. Sixteen journals published two or more included CDW articles (69% of the included CDW articles) (**Figure 3**). *Studies in Health Technology and Informatics* (30 included CDW articles) also publishes conference papers from multiple informatics meetings. The six publications with the most CDW articles (68/137) are informatics journals, as are several others.

From a geographical perspective, North America and Europe lead with the highest contributions in articles numbers, having 61 and 53 respectively. They are followed by Asia with 13, Australia with 5, South America with 4, and Africa with 1. Notably, eight research groups have contributed more than two articles each. These groups include those led by Frank Puppe, James J. Cimino, Raphael W. Majeed, Andrew Post, Shawn Murphy, Griffin M. Weber, Michael Marschollek, and Michael G. Kahn.

Article Type

Case studies comprised the most articles, at 122 case study articles, followed by 5 qualitative study articles, 2 systematic literature review articles, 1 survey article, and 7 discussion articles.

CDW Focus

Over half of the articles reported CDWs that were designed to support or supported research only (**Table 2**). An additional 12% of the articles supported research as well as clinical operations. This result could be caused by the bias of the query.

Table 2: Counts and percentages of CDW focus, architecture, and data model.

Theme/Sub-Theme	Category	Number	Percentage	Articles
CDW Focus	Clinical Research	82	60%	[3, 14–94]
	Clinical Operation	38	28%	[95–132]
	Both	17	12%	[133–149]
CDW Design: Architecture	Public	116	84%	[14, 15, 17–36, 38–41, 43–58, 60–69, 72–88, 90–95, 97–104, 106–109, 112, 114, 115, 117, 118, 121–129, 132–141, 143–146, 148]
	Commercial	5	4%	[16, 37, 119, 131, 147]
	Not reported	16	12%	[3, 42, 59, 70, 71, 89, 96, 105, 110, 111, 113, 116, 120, 130, 142, 149]
CDW Design: Data Model	CDM	35	26%	[15, 25, 32, 34, 43, 44, 46, 51, 52, 54–57, 60, 62, 64, 65, 73, 76, 88, 91, 92, 99, 107, 115, 118, 120, 122, 123, 128, 132, 133, 139, 140, 148]
	Custom	31	23%	[14, 16–19, 21, 24, 27, 31, 35, 37, 48, 49, 79–81, 86, 90, 94, 95, 97, 98, 100, 101, 106, 126, 134, 137, 138, 141, 146]
	Open Source	2	1%	[39, 41]
	Not reported	69	50%	[3, 20, 22, 23, 26, 28–30, 33, 36, 38, 40, 42, 45, 47, 50, 53, 58, 59, 61, 63, 66–72, 74, 75, 77, 78, 82–85, 87, 89, 93, 96, 102–105, 108–114, 116, 117, 119, 121, 124, 125, 127, 129–131, 135, 136, 142–145, 147, 149]
CDW Design: Data Domain	Longitudinal Clinical Data	89	65%	[14–20, 25, 26, 29–31, 33, 35–37, 40–42, 44–46, 49–51, 53–60, 62–67, 69, 71–73, 76, 79, 80, 83, 86, 88, 89, 91, 92, 95, 97–99, 103–108, 114, 115, 117–121, 123–126, 128–131, 133, 134, 136–139, 142, 144, 145, 147–149]
	Disease Specific Longitudinal Data	23	17%	[21, 23, 24, 27, 32, 34, 43, 47, 61, 74, 77, 78, 81, 84, 85, 87, 90, 93, 94, 100, 102, 135, 141]
	Imaging Data	11	8%	[22, 39, 48, 49, 75, 97, 109, 127, 135, 136, 145]
	Unstructured Clinical Data	11	8%	[19, 29, 31, 38, 62, 112, 115, 119, 132, 135, 144]
	Administrative Data	4	3%	[30, 52, 122, 140]
	Biological Molecular Data	3	2%	[70, 135, 146]
	Blood Bank Data	2	1%	[101, 135]
	Not reported	10	7%	[3, 28, 68, 82, 96, 110, 111, 113, 116, 143]

CDW Design

CDW Architecture

After the initial release of Biomedical Translational Research Information System (BTRIS) in July 2009 [24, 28, 135], such home-built architectures were used publicly as CDW solutions; of 121 articles that reported architecture, only 5 were commercial, eg, INDEPTH [37] and PCRC BIMS database [16]. The studies with the home-built architectures used concepts from enterprise data warehousing customized architecture, eg, Entity-Attribute-Value (EAV) model [14, 17]. Of the public architectures, Informatics for i2b2 was the most popular platform used, reported in 22 articles after 2014. Most CDWs refreshed data by extract-transfer-load (ETL) methods. Only a few studies used a direct copy from the source system method [53] or materialized views [119, 121] for a particular study area to provide a query or report service. Sixteen articles (7 discussions, 4 qualitative studies, 2 case studies, 2 systematic literature reviews, and 1 survey) did not report CDW architecture. These articles did not focus on a specific CDW so architecture was not applicable, or the authors did not mention it.

Data Model

CDMs were widely used (36 used CDMs) to map CDW data for sharing, but were only sometimes used as the data structure itself [122, 139]. Typical CDM for-sharing examples include one developed as an interface to serve data from i2b2 in FHIR format (SMART-on-FHIR interface is the first instance we found that allowed i2b2 sites to provide data access in FHIR format) [118]; and another that used the i2b2 CDM to provide data to the TriNetX research network [73]. Custom data models were designed using different schemas, for example, star [141] or snowflake [106]. Commercial and custom data models were often used for CDW structures, for example, one study used another hospital's published specification from Epic EHR system and Clarity data model [86]. Sixty-nine articles did not report data model information, comprising 56 case studies, 6 discussions, 4 qualitative studies, 2 systematic literature reviews, and 1 survey.

Data Domain

Three articles discussed special longitudinal clinical data differently than others: one included longitudinal clinical data from dental care [139]; the other two [30, 65] focused on population and socioeconomic data. Twenty-three articles (17%) focused on disease-specific longitudinal clinical data collected for specific disease research (**Table 2**). Among these, one collected COVID-19 data with time series data, weather, lockdowns, and other variables [90]. Imaging data and unstructured data were also prevalent data types, each appearing in 11 articles. Also included were administrative data (4), biological molecular data (3), and blood bank data (2). Ten articles did not mention specific data domains, including survey, systematic review, discussion, and conference workshop articles.

Semantic Data Representation

Thirty-nine articles discussed or mentioned semantic data representation methods used in the CDW. Twenty-four of these reported use of ontologies in CDWs, such as i2b2, and protégé, which is an HL7 FHIR-driven ontology to map with OMOP CDM or additional custom developed ontologies in data sources. In 19 articles, CDWs used existing standards, including ICD-9, ICD-10, SNOMED, Unified Medical Language System (UMLS), and HL7 data sharing standards. Five of the articles combined both ontology and standards-based terminologies in data sources to fit their requirements for CDWs [14, 28, 35, 125, 149]. Only one study built custom vocabulary for intensive care unit (ICU) data [124]. One systematic review focused on ontologies used in CDWs [146].

Work/Improvement Foci

Data integration was the most frequent improvement focus followed by CDW implementation, data querying, data quality and uses of data (**Table 3**).

CDW Governance

Governance Structure

Twenty-two articles reported on CDW governance structures [25, 27, 29, 31, 32, 34, 37, 41, 43, 45, 50, 55, 71, 85, 87, 97, 103, 105–107, 114, 139]. One of these articles mentioned the existing institutional data governance structure from CTSA [55]. Ten discussed in varying detail their warehouse-specific data governance structure for data access, data definition, data release, data quality management, and data warehouse procedures [31, 32, 34, 37, 43, 45, 50, 85, 87, 114] (see **Table 4**). Nine discussed leveraging extant EHR data governance structure for CDW data definition [103], CDW data access [25, 27, 29, 139] and/or CDW data quality management [25, 27, 41, 97, 106, 107]. One article described data governance challenges and possible solutions across multiple CDWs [105]. One article examined CDW practices and issues at CTSA hubs including data governance [71]. The reported governance foci are shown in **Table 4**.

End-user Support and Documentation

Forty-five articles mentioned end-user support or documentation, most focusing on support tools. Forty of them discussed developing an end-user interface to provide data querying, data reporting, and data visualization capabilities. Among these, 11 articles indicated that the interface built was based on an i2b2 web client [16, 21, 25, 41, 44, 46, 50, 55, 91, 107, 122]. Four articles discussed end-user functionality to generate data reports [48, 98, 112, 141]. One article discussed the provision of a support team to assist end-user with the extraction and documentation of datasets [37].

Staff Training

Seven articles mentioned staff training requirements. Two mentioned that they provided training through a series of programs and workshops [37, 114]. Another two provided training on privacy and security for all staff [36, 135]. The latter would be done in all HCOs today in support of HIPAA compliance and lack of mention may have been due to its assumed presence. In two of the seven articles, a council was formed to address a training need [36, 85]. Two articles mentioned staff training challenges and requirements [71, 105].

Table 3: Counts and percentages for categories of work/improvement foci.

Category	Description/Type	Number	Percentage	References
Data Integration	Mapped and loaded one or multiple data sources to the CDW	45	33%	[14–17, 22, 28, 30, 37, 40, 41, 43–47, 55, 57, 63, 64, 73, 75, 79–81, 85–88, 90–93, 98, 99, 102, 109, 110, 124, 126, 129, 131, 134, 135, 141, 142]
	Added Health Information Exchange (HIE) data	5	4%	[35, 39, 95, 106, 107]
	Extracted data from report documents and loaded into the CDW	5	4%	[61, 84, 112, 114, 132]
	Combined manually entered data and electronically imported data	4	3%	[21, 23, 78, 136]
	Linked to external public data (ie, Socioeconomic Index)	1	1%	[65]
	Collected via portable device	1	1%	[94]
CDW Implementation	The articles described the process of building a data warehouse, including but not limited to gather requirements, create infrastructure, choose a data model, connect to data sources, data import, data governance and data analysis	43	31%	[18, 27, 29, 31–34, 39, 41, 46, 49–51, 69, 70, 76–79, 81, 82, 85–88, 90, 93, 94, 104, 105, 108, 109, 115, 117, 122, 123, 125, 126, 138, 140, 141, 145, 149]
Data Querying	The articles described pulling requested data from CDW for projects	25	18%	[26, 31, 43, 47, 48, 53, 54, 57, 60, 62–64, 74, 76, 83, 91, 104, 108, 118–121, 125, 127, 148]
Data Quality	The articles described the process of scientifically and statistically evaluating data to determine whether they meet the quality required for projects and are of the right type and quantity to be able to support their intended use	13	9%	[20, 32, 34, 58, 59, 66, 67, 89, 98, 105, 134, 137, 147]
Uses of Data (Non-clinical decision making)	The articles described the purpose for which the data were used	10	7%	[3, 29, 30, 36, 42, 66, 69, 71, 101, 146]
Standardization for Sharing	The articles described the process of bringing clinical data into a common format that allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies	8	6%	[52, 68, 80, 96, 99, 111, 128, 133]
Data Visualization	The articles described a tool that provides an accessible way to see and understand trends, outliers, and patterns in clinical data by using visual elements such as charts, graphs, and maps	7	5%	[24, 30, 56, 62, 72, 101, 146]
Clinical Decision Making	The articles described a health information technology to provide clinicians, staff, patients, or other individuals with data, knowledge, and person-specific information for health and health care	6	4%	[24, 30, 97, 103, 113, 116]
Data Acquisition	The articles described the process of collecting data for a clinical project	5	4%	[19, 21, 36, 41, 69]
Protected Health Information (PHI)	The articles described the provision of data policy rules built into data warehouses to protect health information safety and how hospitals and other healthcare providers using and sharing protected health information	5	4%	[31, 83, 100, 131, 136]

Financial Sustainability

One article mentioned that they addressed costs and sustainability and needed to raise revenue to support future work [138]. Another article discussed financial sources and sustainability as issues and summarized CDWs financial sources [71].

Discussion

The review highlights areas of CDW focus in the literature, and exemplar projects, with brief explanation of key CDW concepts. Given the necessity of data for health care quality improvement and performance measurement since the landmark 1999 report To Err Is Human: Building

Table 4: Governance foci reported.

Governance foci	Discussed in detail	Mentioned in brief
Data Definition: decision-making surrounding data to be included in the CDW or tools for end-users explaining in the CDW	[45, 103]	–
Data Warehouse Procedures: including shared institutional decision-making surrounding ETL processes, data de-identification, and other standard operating procedures	[31, 43, 45, 50, 55, 85, 114]	–
Data Access: decision making surrounding end-users' access policies	[25, 27, 29, 31, 32, 43, 50, 87, 114, 139]	[37]
Data Release: decision-making surrounding approval required for use of data for specific projects	[43, 50, 114]	–
Quality Management: Oversight quality assurance and decision-making surrounding improving the quality of data	[25, 27, 31, 32, 34, 41, 55, 85, 97, 114]	[43, 106, 107]

a Safer Health System, and for clinical research since the CTSA program, the lack of emphasis on data warehousing is striking. Several factors may be at play. The international Good Clinical Practice (GCP) standard for clinical research did not emphasize data handling until the 2018 revision. Similarly, a recent systematic review of Data Management Plan requirements found an overwhelming emphasis on data sharing with little attention devoted to data collection and processing [150]. Others have attributed this to a lingering perception of data collection and handling as clerical and largely ignore their impact on research results [150, 151]. As evidenced by the literature, the lack of scientific consideration is accompanied by a lack of evidence-based practice recommendations and lack of defined professional competencies, with the latter recently called to attention by the Healthcare Data Analytics Association (HDAA) [152]. These findings, consistent with recent qualitative and survey work within the CTSA organizations, underscore the gap between the touted importance of data to results and the lack of attention and resources necessary to ensure that data are capable of supporting research.

CDW architecture: CDW architecture depends heavily on the specific data domains, eg, intensive care unit, dentistry, or specific disease registries, it is designed to store. No one design fits all needs and the literature bears out variability across the spectrum from, (1) building a data warehouse from scratch – the most frequently reported, (2) customizing an existing architecture for a specific domain, or (3) using a commercial data warehousing product. Since we did not include the terms of disease-specific data registry or commercial architecture in the query, the architecture reported here could under-represent the disease-specific or commercial CDWs. The majority, 88% (121) of the articles reported whether they were commercial or public architectures [153]. The degree to which further details on products or schemas was reported, varied across the literature; for example, one study described the usage of facts and dimensions in building their data warehouse to process data from electronic systems [106, 107] and another reported the usage of existing warehouse architecture using Microsoft products in their implementation [134]. In addition, one article reported the commercial product name INDEPTH for their warehouse but very limited details on the architecture were disclosed [37]. Of the public architectures, the majority used i2b2, which is similar to the star and snowflake schemas [154], followed by domain-specific data repository models that used relational tables [4].

Data model: Some CDWs are built on a generic relational database data model (as discussed in the architecture section). However, many are now constructed using a common data model (CDM) and open-source clinical data frameworks devised for standard representation of health information that provides better data analysis and data exchange. However, the term "common data model" was not included in the query. The data model reported here is only based on the literature as implemented as a CDW. It could therefore under-represent the other CDMs, except i2b2. In our results, 35 studies utilized CDMs that include OMOP, CDISC operation data model, and i2b2 one use of which is as an open-source clinical data model (see **Table 2**). Three studies extended the i2b2 querying mechanism to support data analysis of diverse underlying CDMs, eg, OMOP and PCORnet, which provided a means to separate data model and querying techniques [60, 91, 122]. One utilized an i2b2 application programming interface (API) and extended it to query OMOP and PCORNet data models, with no significant performance degradation in their evaluation [60]. Another highlighted the lack of querying interface for OMOP and developed an algorithm that can translate any given i2b2 query to run against OMOP data model programmatically [91]. That algorithm took advantage of i2b2's webclient and its metadata, which prevented the need for a full i2b2 data model installation. The third used i2b2's new multi-fact table to use OMOP CDM as a source for i2b2 observational data without changes to the i2b2 source code [122]. That use case demonstrated the combination of OMOP CDM, data management of i2b2, and an interface that allows querying of both i2b2 and OMOP. No CDM, however, covers all use cases and needs, so custom in-house and proprietary models specifically designed for the disease domain and type of data are still developed [155]. Two studies implemented a custom data model to capture perioperative anesthesia and omics data [86, 146]S.

Data domain: CDWs are generally built to provide a holistic view of patient care by the inclusion and sometimes the integration of other data. This can include external data, such as claims; molecular data, such as metabolomics; multi-level data, such as viral sequences; environmental measures, such as air and water quality; place-based social measures, such as neighborhood transience; and unstructured data [156]. We found that seven studies [19, 29, 31, 62, 115, 119, 144] developed a system that combined longitudinal clinical data with unstructured documents while four [49, 97, 136, 145] used imaging data along with longitudinal clinical data. In contrast, two studies focused on single clinical specialty for better management and querying [22, 109]. To facilitate clinical and translational research, however, combining different data domains is important, and only one study by Cimino included comprehensive domain representation ranging from imaging, unstructured, molecular, blood bank, and longitudinal clinical data [135]. An area for additional work is developing methodologies for data integration.

Data integration/ETL: Data intake from source systems and their integration into a data warehouse occurs through a process commonly known as ETL. ETL is a sequential process that starts with identifying needed clinical variables from the source systems, includes mapping the data into the warehouse data model, and then implementing computer programs that use the mapping to write the external data to the warehouse tables. Common ETL operations include deidentification, assignment of unique identifiers, reformatting data, recoding data to standard terminologies and ontologies, and data linkage. A similar process is often followed to deliver data from an institution's data warehouse for secondary use; in this case, the warehouse data are often mapped to a CDM that supports pooling data from multiple institutions. ETL processes have received the most attention in the literature. A systematic review focused on ETL and discussed ETL steps in detail [149]. Nineteen additional studies discussed ETL and focused on: terminology mapping [39, 46, 55, 73, 92, 106, 107, 124], data deidentification [129], specific data type extraction [15, 87], unstructured data ETL [61, 84, 109], automatically generated ETL [40, 46, 114, 124], re-use ETL [135], real-time ETL [88, 129], and ETL efficiency [86]. Adding these health care-specific practices to general data warehousing frameworks may provide a foundation for health care-informed ETL best practices in health care data warehousing. Alternatively, others such as a FHIR-to-i2b2 transformation toolkit [123], which directly transforms primary EHR data from standardized FHIR resources into I2b2 CDM, focused on automating ETL processes and decreasing the laborious creation and maintenance of ETL processes.

Data quality: Clinical decision-making and research that relies on data from CDWs require high-quality data. Quality management was an important area for CDW development. However, only 9 per cent (13) of the articles reported data quality assessment (DQA) effort (11 case studies and 2 discussion papers). Rule-based [20, 34, 66, 98, 137] and redundancy-based [20, 58, 134, 147] methods were the most commonly used DQA methods (in 5 and 4 articles, respectively). Crowdsourcing [34], data profiling [32] and distribution method [67] were reported by one article each. Seven articles assessed different data quality dimensions, including completeness [20, 58, 66, 98, 147], validity [20, 32, 58], consistency [66, 98], accuracy [98, 147], plausibility [66, 67] and timeliness [20]. Different DQA processes were identified from nine articles, including routine DQA in the CDW [20, 32, 34, 67, 98, 147], DQA during data extraction [58], comparing data in the CDW with its source data [134], and DQA at data entry [137]. Only two articles reported having processes in place and commitments from source system owners to help to investigate, remediate, or make changes to prevent future occurrence of data quality problems [20, 98].

Semantic data representation: The key to successfully integrate and represent clinical data is through the use of semantic interoperability methods that describe complex health care information in computable ways [157, 158]. Standardized clinical terminologies and ontologies are ways that provide semantic representation of clinical concepts to achieve automated interoperability and to enable the best possible use of such data [159, 160]. However, a challenge is that any one terminology or ontology is not suitable for all purposes. The majority of the studies in our review used or extended i2b2 ontology in their work because of its simple, flexible, open-source data representation [161]. Two studies [15, 33] adopted the Protégé ontology framework for building a custom ontology whereas one study [135] developed a custom tool to address the practical needs of real-world data and to address the limitations of the i2b2 ontology eg, support for multiple hierarchies. Use of standard terminologies is vital to facilitate data sharing for clinical research and automated clinical decision support [162]. The majority of the studies used ICD-9 and ICD-10 standard terminologies for concept mapping. Systemized Nomenclature of Medicine – Clinical Terms (SNOMED-CT), Logical Observation Identifiers Names and Codes (LOINC), and Unified Medical Language System (UMLS) were also used frequently. One study [133] used ICD-9, SNOMED, UMLS, and LOINC for their mapping needs, and two studies used both ICD and CPT [29, 32]. To fully exploit the potential of both standard terminologies and ontologies, five studies developed custom ontologies that incorporated standard terminologies such as ICD, UMLS, and SNOMED [14, 28, 35, 125, 149]. One study described the limitations of terminology mapping tools in Intensive Care Units (ICU) and created a concept vocabulary of 942 clinical parameters to link concepts during the COVID-19 emergency [87]. Additional work is needed to build easy-to-use flexible terminology mapping tools for streamlined concept mapping in ICU settings. Four studies used document-based clinical document architecture (CDA) and other HL7 data standards [19, 35, 106, 107] and one study [57] used Clinical Data Interchange Standards Consortium (CDISC) data standards rather than terminologies and ontologies for knowledge representation and sharing. A likely future direction for CDWs is more HL7 FHIR data standards-based representation work for interoperability.

CDW governance structure: Among the articles that mentioned governance structure, two main types of structure were identified, including extant EHR governance (focus on data itself) and warehouse-specific governance (including data management and data warehouse procedures) with nine and 10 articles reported, respectively. No articles after 2015 referred to extant EHR governance structures, but five reported warehouse-specific ones, which is a more comprehensive and specific solution to satisfy data-use related regulatory requirements.

End-user support and documentation: CDW end-users are not familiar with the technical aspects of their projects and lack of data literacy, which is a challenge for CDW staff [105]. Many articles discussed query interfaces or tools to support end-users. However, only two articles mentioned the need to improve end-users' data literacy with respect to knowledge about clinical data, use of EHR data in research, and protection of patient privacy, which highlights an area for more work [71, 105].

Staff training: Although two articles mentioned that they provided staff training through a series of programs and workshops to support data literacy [37, 114], none reported needs of assessment results, position-specific training requirements, or training plan details. To our knowledge, there are no publicly shared training modules for health care data warehousing professionals. HDAA has recently undertaken a campaign to identify professional competencies for healthcare data warehousing [152].

Financial sustainability: Only two (2/137) of the articles mentioned CDW cost models or sustainability [71, 138]. However, given that 48 articles mentioned grant funding, usually in the Acknowledgements section, it is likely that CDWs lack long-term funding and sustainability. One article discussed two approaches that enterprise data warehouses for research use to fund themselves: fee-for-service and full-time equivalent funding, but the article notes that the ability to fulfill requests does not keep up with demand, which indicates financial and staffing deficits given demand [71]. In light of the cost pressures in AMCs, a shift in CTSA emphasis, from clinical and translational research (CTR) support to clinical and translational science (CTS), and the growing need to provide data to support the development of clinical researchers and the sustainability of clinical research programs needs to be prioritized. In addition, information on the real costs and benefits of data warehousing to institutions is drastically and urgently needed.

Geographical differences: We categorized and analyzed the articles based on geographical distribution. North America and Europe lead the research in CDWs, though their emphases differ as a result of variations in health care systems, research focus, regulatory environments, and data infrastructure. In North America, CDW research concentrates on integrating CDWs within large, complex health care delivery systems, including hospital networks, insurance companies, and research institutions. Many CDWs are developed by individual health care organizations or through collaborations between academic and private sectors. In contrast, European CDW research places emphasis on national or regional health care systems, aiming for unified data integration across public health networks. This approach is often supported by government-led initiatives or by large-scale public health projects, fostering uniform standards and interoperability across countries. This leads to differences in innovation. In North America, there is a high emphasis on leveraging advanced analytics, artificial intelligence (AI), and machine learning to derive insights from clinical data. In contrast, Europe has a strong focus on public health informatics, population health management, and the integration of social determinants of health into their CDW research and applications.

Diversity of contributing groups: The diversity of contributing research groups is notable in the field, as evidenced by the author distribution across 137 articles. Eight key research groups stand out, each contributing three to five articles. In addition, numerous other groups have provided valuable insights, contributing fewer than two articles each. This distribution highlights the essential need for cross-validation and collaborative efforts to enhance the robustness and generalizability of findings across different contexts and settings. Such collaboration among diverse research groups is crucial for advancing the field.

Gap analysis: We reviewed 2021 to 2022 HDAA conference presentations and 2023 HDAA conference submitter-identified tags to distill 30 current topics at the applied practitioner forefront [2]. The 50 presentations in 2021 and 2022 conferences covered 26 topics. Among these, predictive analytics (17), machine learning (8) and staffing (4) were the top 3 topics. Data lakes, data fabric, Microsoft Azure Cognitive Service and ChatGPT-4 were added in 2023 as new tags. We matched these topical categories to our included articles. Articles could be matched to more than one category. **Figure 4** shows where the published literature aligns with topics from HDAA and where there are particular publication gaps. Thirteen topics were covered by at least 10 articles, and 17 topics by less than 10, of which, equity, work-life balance, career development, and specific data technologies or services (eg, data fabric, data mesh, data lakes, Microsoft Azure Cognitive Service, ChatGPT-4) were not covered by any articles. Some topics have minimal or non-published articles due to the emphasis of the query, such as data lakes and Microsoft Azure Cognitive Service. However, these topics should be discussed in the literature as do similar articles involving technical developments and CDW-related sociotechnical issues related to staff, end-users, and decision-makers. To reach a state of health care data analytics maturity, more work and published dissemination on data quality improvement and approaches for predictive analytics are needed [163].

Lastly, the research methods found in the literature were largely single site demonstrations or opinion articles based on experiences at a single site. To wit, all were one site/project case studies, discussions, or reviews except for six articles: five qualitative studies and a survey. Four of the five covered single aspects: two single-site/project articles [70, 100] and two multi-site articles [110, 116]. Only two were multi-site, multi-aspect studies, a survey [3]

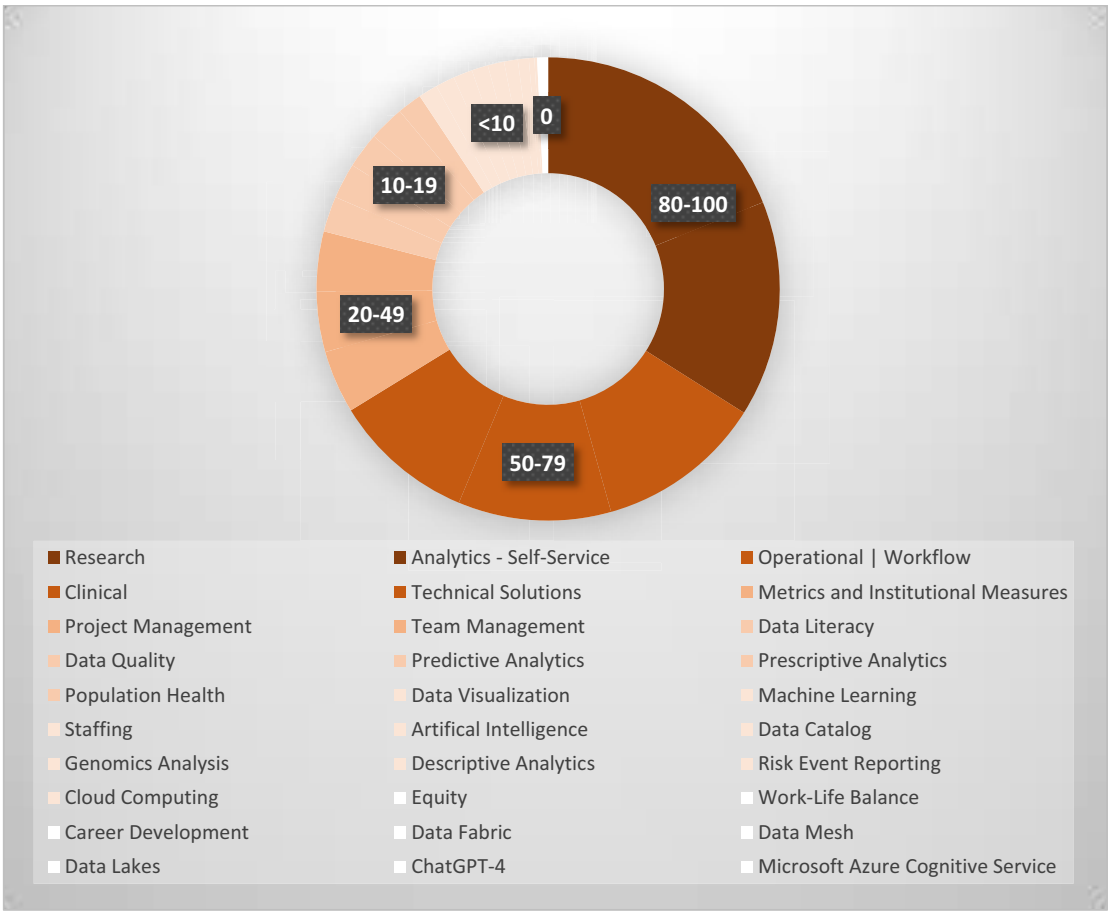

Figure 4: Gap assessment between what the CDW literature addresses and topics that need more publication to advance knowledge-sharing and CDW maturity.

and one qualitative study [71]. To increase the evidence-base for CDW practice, more multi-site, multi-aspect CDW benchmarking and reporting are needed.

In recent years, specific requirements for CDWs have emerged, driven by the unique needs of clinical research and the growing complexity of data. Best practices specific to CDWs have been evolving to meet the challenges of managing clinical research data. These practices include: (1) data integration and interoperability to facilitate seamless integration of diverse clinical data sources and ensure interoperability within the CDW; (2) data governance to ensure data integrity, confidentiality, and compliance with regulatory requirements within the CDW environment; (3) data quality to implement procedures to ensure high-quality, accurate, and reliable clinical data within the CDW; and (4) metadata management to maintain comprehensive metadata to support data discoverability, provenance tracking, and data lineage within the CDW.

Despite the established best practices and newly articulated requirements, significant gaps remain in the CDW literature that need to be addressed to enhance the efficacy of clinical research. The gap analysis revealed the following issues. First, while general data integration practices are well-defined, there is a lack of detailed strategies tailored to integrating diverse clinical data specific to research needs. Second, current data quality practices may not sufficiently address the nuances of clinical research data, such as the need for high accuracy in longitudinal studies. Third, many CDWs are not designed to scale efficiently with the rapidly growing data volumes and evolving research methodologies. Fourth, the lack of robust metadata and provenance tracking systems can hinder data discoverability and reproducibility in clinical research. Last, ethical considerations and patient consent protocols are not always integrated seamlessly into data management practices. By addressing these gaps and by implementing the recommended best practices, we can significantly enhance the utility and reliability of CDWs in clinical research settings, ultimately leading to more robust and impactful research outcomes.

Limitations: There are several limitations in this review that we wish to acknowledge. First, we only searched the query in PubMed as it is a comprehensive resource in the clinical arena. We could retrieve more related literature with other search engines. Second, we included the terms of medical record in the query to cover more related literatures. However, this process might miss the articles that did not mention the terms of medical record and get noisy articles. Beside medical records, the specific data registry, such as cancer, cardiovascular, and COVID data, are important considerations, but it was beyond the scope of this review. Adding terms of specific disease registry would return more CDW articles. Third, we emphasized the literature as implemented as a CDW, especially academic-implemented CDWs, due to only including the terms of i2b2 in the query. Doing this led to an over-representation

of i2b2 use and an under-representation other CDMs, commercial CDWs, and specific methodologies, such as data model and data quality. This bias could be corrected by including the terms of CDM, commercial CDWs, OMOP or OHDSI (Observational Health Data Sciences and Informatics), or by expanding the review to include EHR-published papers (such as EHRN.org). Last, we did the final query run in September 2021. However, an explosion of publications related to CDW happened after 2021 as a result of the COVID pandemic. The query needs to be modified and re-run to get more recent and comprehensive information.

Conclusions

The study summarizes the studies from 2011 to 2021 with important topics in CDW. It indicates the topics that have been significantly developed and the aspects that need additional focus and reporting in CDW between existing general data management best practices and recently articulated requirements for research data. More multi-site and multi-aspect studies are needed to foster maturity at CDWs. The findings of this scoping review will help readers to understand current CDW methodologies and improvement foci and we hope will inspire HCOs to invest in CDW deployment, ongoing optimization for expanded utility of existing CDWs, and increased CDW benchmarking and knowledge-sharing.

Summary table

What is known about this problem:

1. Significant efforts have been made by HCOs to develop CDWs.
2. CDWs have proved their value in clinical decision making and clinical research secondary data use.

What this article contributes:

1. The article reviews and summarizes the CDW studies from 2011 to 2021 based on CDW focus; architecture type; data model; data domain; work and improvement foci; semantic data representation; governance structure; end-user support tools and documentation; staff training; financial sources; and sustainability.
2. The article indicates the topics that have been significantly developed and the aspects that need additional focus and reporting in CDW.
3. The findings should inspire HCOs to invest in CDW deployment, ongoing optimization for expanded utility of existing CDWs, and increased CDW benchmarking and knowledge-sharing.

Additional Files

The additional files for this article can be found as follows:

- **Supplementary File 1.** Appendix 1. DOI: https://doi.org/10.47912/jscdm.320.s1
- **Supplementary File 2.** Appendix 2. DOI: https://doi.org/10.47912/jscdm.320.s2

Acknowledgements

We would like to thank Cynthia C. Caton, medical librarian, University of Arkansas for Medical Sciences, for her expert collaboration in formulating our PubMed query.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

ZW, MS, SS, and CKC contributed to all aspects. MNZ contributed to study design, PubMed query construction, article abstraction/categorization categories. MG contributed to the study design and query construction, with significant effort on the initial version of this project. ES contributed to literature abstraction on the initial version of this project. All authors conducted the final review.

References

1. **Norton SL, Buchanan AV, Rossmann DL, Chakraborty R, Weiss KM.** Data entry errors in an on-line operation. *Comput Biomed Res.* 1981; 14(2): 179–98. DOI: https://doi.org/10.1016/0010-4809(81)90035-5
2. **Healthcare Data & Analytics Association.** [cited 2022 Jul 29]. Available from: https://www.hdwa.org/.
3. **MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N.** Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. *J Am Med Inform Assoc.* 2012; 19(e1): e119–24. DOI: https://doi.org/10.1136/amiajnl-2011-000508
4. **Inmon WH.** Building the data warehouse. John Wiley & Sons; 2005.
5. Scoping reviews: what they are and how you can do them: The Cochrane Collaboration; [cited 2022 Oct 4]. Available from: https://training.cochrane.org/resource/scoping-reviews-what-they-are-and-how-you-can-do-them.
6. **Canadian Institutes of Health Research.** A Guide to Knowledge Synthesis.; [cited 2022 Oct 4]. Available from: https://cihr-irsc.gc.ca/e/41382.html.
7. **Arksey H, O'Malley L.** Scoping studies: towards a methodological framework. International *Journal of Social Research Methodology*. 2005; 8(1): 19–32. DOI: https://doi.org/10.1080/1364557032000119616
8. **Levac D, Colquhoun H, O'Brien KK.** Scoping studies: advancing the methodology. *Implementation Science.* 2010; 5(1): 1–9. DOI: https://doi.org/10.1186/1748-5908-5-69
9. **National Institutes of Health.** National Library of Medicine. National Center for Biotechnology Information. [cited 2022 Oct 4]. Available from: https://pubmed.ncbi.nlm.nih.gov/.
10. **Fiorini N, Lipman DJ, Lu Z.** Towards PubMed 2.0. *Elife.* 2017; 6. DOI: https://doi.org/10.7554/eLife.28801
11. EPIC system 2024. Available from: https://www.epic.com/.

12. **Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD,** et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ.* 2021; 372: n160. DOI: https://doi.org/10.1136/bmj.n160
13. **Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D,** et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine.* 2018; 169(7): 467–73. DOI: https://doi.org/10.7326/M18-0850
14. **Hu H, Correll M, Kvecher L, Osmond M, Clark J, Bekhash A,** et al. DW4TR: A data warehouse for translational research. *J Biomed Inform.* 2011; 44(6): 1004–19. DOI: https://doi.org/10.1016/j.jbi.2011.08.003
15. **Post A, Kurc T, Overcash M, Cantrell D, Morris T, Eckerson K,** et al. A temporal abstraction-based extract, transform and load process for creating registry databases for research. *AMIA Jt Summits Transl Sci Proc.* 2011; 2011: 46–50.
16. **Späth MB, Grimson J.** Applying the archetype approach to the database of a biobank information management system. *Int J Med Inform.* 2011; 80(3): 205–26. DOI: https://doi.org/10.1016/j.ijmedinf.2010.11.002
17. **Wade TD, Hum RC, Murphy JR.** A dimensional bus model for integrating clinical and research data. *J Am Med Inform Assoc.* 2011; 18 Suppl 1(Suppl 1): i96–102. DOI: https://doi.org/10.1136/amiajnl-2011-000339
18. **de Mul M, Alons P, van der Velde P, Konings I, Bakker J, Hazelzet J.** Development of a clinical data warehouse from an intensive care clinical information system. *Comput Methods Programs Biomed.* 2012; 105(1): 22–30. DOI: https://doi.org/10.1016/j.cmpb.2010.07.002
19. **Li Z, Wen J, Zhang X, Wu C, Li Z, Liu L.** ClinData Express–a metadata driven clinical research data management system for secondary use of clinical data. *AMIA Annu Symp Proc.* 2012; 2012: 552–7.
20. **Byrd JB, Vigen R, Plomondon ME, Rumsfeld JS, Box TL, Fihn SD,** et al. Data quality of an electronic health record tool to support VA cardiac catheterization laboratory quality improvement: the VA Clinical Assessment, Reporting, and Tracking System for Cath Labs (CART) program. *Am Heart J.* 2013; 165(3): 434–40. DOI: https://doi.org/10.1016/j.ahj.2012.12.009
21. **Choi IY, Park S, Park B, Chung BH, Kim CS, Lee HM,** et al. Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system. *Prostate Int.* 2013; 1(2): 59–64. DOI: https://doi.org/10.12954/PI.12015
22. **Franke T, Gruetz R, Dickmann F.** Functional requirements for a central research imaging data repository. *Stud Health Technol Inform.* 2013; 192: 298–302.
23. **Hong MK, Yao HH, Pedersen JS, Peters JS, Costello AJ, Murphy DG,** et al. Error rates in a clinical data repository: lessons from the transition to electronic data transfer–a descriptive study. *BMJ Open.* 2013; 3(5). DOI: https://doi.org/10.1136/bmjopen-2012-002406
24. **Manning JD, Marciano BE, Cimino JJ.** Visualizing the data – using lifelines2 to gain insights from data drawn from a clinical data repository. *AMIA Jt Summits Transl Sci Proc.* 2013; 2013: 168–72.
25. **Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W,** et al. The analytic information warehouse (AIW): a platform for analytics using electronic health record data. *J Biomed Inform.* 2013; 46(3): 410–24. DOI: https://doi.org/10.1016/j.jbi.2013.01.005
26. **Weber GM.** Federated queries of clinical data repositories: the sum of the parts does not equal the whole. *J Am Med Inform Assoc.* 2013; 20(e1): e155–61. DOI: https://doi.org/10.1136/amiajnl-2012-001299
27. **Ardini MA, Pan H, Qin Y, Cooley PC.** Sample and data sharing: observations from a central data repository. *Clin Biochem.* 2014; 47(4–5): 252–7. DOI: https://doi.org/10.1016/j.clinbiochem.2013.11.014
28. **Cimino JJ, Remennick L.** Adapting a Clinical Data Repository to ICD-10-CM through the use of a Terminology Repository. *AMIA Annu Symp Proc.* 2014; 2014: 405–13.
29. **Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S,** et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform.* 2014; 52: 28–35. DOI: https://doi.org/10.1016/j.jbi.2014.02.003
30. **Haque W, Urquhart B, Berg E, Dhanoa R.** Using business intelligence to analyze and share health system infrastructure data in a rural health authority. *JMIR Med Inform.* 2014; 2(2): e16. DOI: https://doi.org/10.2196/medinform.3590
31. **Horvath MM, Rusincovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM.** Modular design, application architecture, and usage of a self-service model for enterprise data delivery: the Duke Enterprise Data Unified Content Explorer (DEDUCE). *J Biomed Inform.* 2014; 52: 231–42. DOI: https://doi.org/10.1016/j.jbi.2014.07.006
32. **Laws R, Gillespie S, Puro J, Van Rompaey S, Quach T, Carroll J,** et al. The Community Health Applied Research Network (CHARN) Ddata warehouse: a resource for patient-centered outcomes research and quality improvement in underserved, safety net populations. *eGEMS (Wash DC).* 2014; 2(3): 1097. DOI: https://doi.org/10.13063/2327-9214.1097
33. **Lozano-Rubí R, Pastor X, Lozano E.** OWLing clinical data repositories with the ontology web language. *JMIR Med Inform.* 2014; 2(2): e14. DOI: https://doi.org/10.2196/medinform.3023
34. **Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G,** et al. The HMO research network virtual data warehouse: a public data model to support

collaboration. *eGEMS (Wash DC).* 2014; 2(1): 1049. DOI: https://doi.org/10.13063/2327-9214.1049

35. **Wang Y, Pakhomov S, Dale JL, Chen ES, Melton GB.** Application of HL7/LOINC document ontology to a university-affiliated integrated health system research clinical data repository. *AMIA Jt Summits Transl Sci Proc.* 2014; 2014: 230–4.
36. **Ghany A, Keshavjee K.** A platform to collect structured data from multiple EMRs. *Stud Health Technol Inform.* 2015; 208: 142–6.
37. **Herbst K, Juvekar S, Bhattacharjee T, Bangha M, Patharia N, Tei T,** et al. The INDEPTH data repository: an international resource for longitudinal population and health data from health and demographic surveillance systems. *J Empir Res Hum Res Ethics.* 2015; 10(3): 324–33. DOI: https://doi.org/10.1177/1556264615594600
38. **Li D, Rastegar Mojarad M, Li Y, Sohn S, Mehrabi S, Komandur Elayavilli R,** et al. A frequency-based strategy of obtaining sentences from clinical data repository for crowdsourcing. *Stud Health Technol Inform.* 2015; 216: 1033–4. DOI: https://doi.org/10.1145/2808719.2808752
39. **Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG.** Archetype-based data warehouse environment to enable the reuse of electronic health record data. *Int J Med Inform.* 2015; 84(9): 702–14. DOI: https://doi.org/10.1016/j.ijmedinf.2015.05.016
40. **Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch HU, Bürkle T,** et al. Ontology-based data integration between clinical and research systems. *PLoS One.* 2015; 10(1): e0116656. DOI: https://doi.org/10.1371/journal.pone.0116656
41. **Pittman CA, Miranpuri AS.** Neurosurgery clinical registry data collection utilizing Informatics for integrating biology and the bedside and electronic health records at the University of Rochester. *Neurosurg Focus.* 2015; 39(6): E16. DOI: https://doi.org/10.3171/2015.9.FOCUS15382
42. **Price LE, Shea K, Gephart S.** The Veterans Affairs's corporate data warehouse: uses and implications for nursing research and practice. *Nurs Adm Q.* 2015; 39(4): 311–8. DOI: https://doi.org/10.1097/NAQ.0000000000000118
43. **Weber GM.** Federated queries of clinical data repositories: scaling to a national network. *J Biomed Inform.* 2015; 55: 231–6. DOI: https://doi.org/10.1016/j.jbi.2015.04.012
44. **Bauer CR, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M,** et al. Integrated Data Repository Toolkit (IDRT). A suite of programs to facilitate health analytics on heterogeneous medical data. *Methods Inf Med.* 2016; 55(2): 125–35. DOI: https://doi.org/10.3414/ME15-01-0082
45. **Chelico JD, Wilcox AB, Vawdrey DK, Kuperman GJ.** Designing a clinical data warehouse architecture to support quality improvement initiatives. *AMIA Annu Symp Proc.* 2016; 2016: 381–90.
46. **Haarbrandt B, Tute E, Marschollek M.** Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. *J Biomed Inform.* 2016; 63: 277–94. DOI: https://doi.org/10.1016/j.jbi.2016.08.007
47. **Kaspar M, Ertl M, Fette G, Dietrich G, Toepfer M, Angermann C,** et al. Data linkage from clinical to study databases via an R Data warehouse user interface. Experiences from a large clinical follow-up study. *Methods Inf Med.* 2016; 55(4): 381–6. DOI: https://doi.org/10.3414/ME15-02-0015
48. **Langer SG.** DICOM data warehouse: part 2. *J Digit Imaging.* 2016; 29(3): 309–13. DOI: https://doi.org/10.1007/s10278-015-9830-4
49. **Min L, Liu J, Lu X, Duan H, Qiao Q.** An implementation of clinical data repository with openEHR approach: from data modeling to architecture. *Stud Health Technol Inform.* 2016; 227: 100–5.
50. **Turley CB, Obeid J, Larsen R, Fryar KM, Lenert L, Bjorn A,** et al. Leveraging a statewide clinical data warehouse to expand boundaries of the learning health system. *eGEMS (Wash DC).* 2016; 4(1): 1245. DOI: https://doi.org/10.13063/2327-9214.1245
51. **Boussadi A, Zapletal E.** A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. *BMC Med Inform Decis Mak.* 2017; 17(1): 120. DOI: https://doi.org/10.1186/s12911-017-0513-6
52. **Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR.** A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. *Stud Health Technol Inform.* 2017; 245: 887–91.
53. **Kim HS, Kim H, Jeong YJ, Kim TM, Yang SJ, Baik SJ,** et al. Development of clinical data mart of HMG-CoA reductase inhibitor for varied clinical research. *Endocrinol Metab (Seoul).* 2017; 32(1): 90–8. DOI: https://doi.org/10.3803/EnM.2017.32.1.90
54. **Majeed RW, Stöhr MR, Thiemann VS, Röhrig R, Günther A.** Asynchronous query distribution between multiple i2b2 research data warehouses: Li2b2-SHRINE. *Stud Health Technol Inform.* 2017; 245: 1276.
55. **Post AR, Ai M, Kalsanka Pai A, Overcash M, Stephens DS.** Architecting the data loading process for an i2b2 research data warehouse: full reload versus incremental updating. *AMIA Annu Symp Proc.* 2017; 2017: 1411–20.
56. **Thiemann VS, Xu T, Röhrig R, Majeed RW.** Automated report generation for research data repositories: from i2b2 to PDF. *Stud Health Technol Inform.* 2017; 245: 1289.
57. **Yamamoto K, Ota K, Akiya I, Shintani A.** A pragmatic method for transforming clinical research data from the research electronic data capture "REDCap" to Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM): development and evaluation of

REDCap2SDTM. *J Biomed Inform.* 2017; 70: 65–76. DOI: https://doi.org/10.1016/j.jbi.2017.05.003

58. **Brundin-Mather R, Soo A, Zuege DJ, Niven DJ, Fiest K, Doig CJ,** et al. Secondary EMR data for quality improvement and research: A comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *J Crit Care.* 2018; 47: 295–301. DOI: https://doi.org/10.1016/j.jcrc.2018.07.021
59. **Huser V, Kahn MG, Brown JS, Gouripeddi R.** Methods for examining data quality in healthcare integrated data repositories. *Pac Symp Biocomput.* 2018; 23: 628–33. DOI: https://doi.org/10.1142/9789813235533_0059
60. **Klann JG, Phillips LC, Herrick C, Joss MAH, Wagholikar KB, Murphy SN.** Web services for data warehouses: OMOP and PCORnet on i2b2. *J Am Med Inform Assoc.* 2018; 25(10): 1331–8. DOI: https://doi.org/10.1093/jamia/ocy093
61. **Linkov F, Silverstein JC, Davis M, Crocker B, Hao D, Schneider A,** et al. Integration of cancer registry data into the text information extraction system: leveraging the structured data import tool. *J Pathol Inform.* 2018; 9: 47. DOI: https://doi.org/10.4103/jpi.jpi_38_18
62. **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: 1445–6.
63. **Dietrich G, Krebs J, Liman L, Fette G, Ertl M, Kaspar M,** et al. Replicating medication trend studies using ad hoc information extraction in a clinical data warehouse. *BMC Med Inform Decis Mak.* 2019; 19(1): 15. DOI: https://doi.org/10.1186/s12911-018-0729-0
64. **Gabetta M, Mirabelli M, Klersy C, Musella V, Rizzo G, Pedrazzoli P,** et al. An extension of the i2b2 data warehouse to support REDCap dynamic data pull. *Stud Health Technol Inform.* 2019; 258: 21–5.
65. **Gardner BJ, Pedersen JG, Campbell ME, McClay JC.** Incorporating a location-based socioeconomic index into a de-identified i2b2 clinical data warehouse. *J Am Med Inform Assoc.* 2019; 26(4): 286–93. DOI: https://doi.org/10.1093/jamia/ocy172
66. **Henley-Smith S, Boyle D, Gray K.** Improving a secondary use health data warehouse: proposing a multi-level data quality framework. *eGEMS (Wash DC).* 2019; 7(1): 38. DOI: https://doi.org/10.5334/egems.298
67. **Looten V, Kong Win Chang L, Neuraz A, Landau-Loriot MA, Vedie B, Paul JL,** et al. What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse. *Comput Methods Programs Biomed.* 2019; 181: 104825. DOI: https://doi.org/10.1016/j.cmpb.2018.12.030
68. **Xafis V, Labude MK.** Openness in big data and data repositories: the application of an ethics framework for big data in health and research. *Asian Bioeth Rev.* 2019; 11(3): 255–73. DOI: https://doi.org/10.1007/s41649-019-00097-z
69. **Antonio MG, Schick-Makaroff K, Doiron JM, Sheilds L, White L, Molzahn A.** Qualitative data management and analysis within a data repository. *West J Nurs Res.* 2020; 42(8): 640–8. DOI: https://doi.org/10.1177/0193945919881706
70. **Broekstra R, Aris-Meijer J, Maeckelberghe E, Stolk R, Otten S.** Trust in centralized large-scale data repository: a qualitative analysis. *J Empir Res Hum Res Ethics.* 2020; 15(4): 365–78. DOI: https://doi.org/10.1177/1556264619888365
71. **Campion TR, Craven CK, Dorr DA, Knosp BM.** Understanding enterprise data warehouses to support clinical and translational research. *J Am Med Inform Assoc.* 2020; 27(9): 1352–8. DOI: https://doi.org/10.1093/jamia/ocaa089
72. **Cancé C, Madiot PE, Lenne C, Artemova S, Cohard B, Bodin M,** et al. Cohort creation and visualization using graph model in the PREDIMED health data warehouse. *Stud Health Technol Inform.* 2020; 270: 108–12.
73. **González L, Pérez-Rey D, Alonso E, Hernández G, Serrano P, Pedrera M,** et al. Building an i2b2-based population repository for clinical research. *Stud Health Technol Inform.* 2020; 270: 78–82.
74. **Greenwood AK, Montgomery KS, Kauer N, Woo KH, Leanza ZJ, Poehlman WL,** et al. The AD knowledge portal: a repository for multi-omic data on Alzheimer's disease and aging. *Curr Protoc Hum Genet.* 2020; 108(1): e105. DOI: https://doi.org/10.1002/cphg.105
75. **Kaspar M, Liman L, Ertl M, Fette G, Seidlmayer LK, Schreiber L,** et al. Unlocking the PACS DICOM domain for its use in clinical research data warehouses. *J Digit Imaging.* 2020; 33(4): 1016–25. DOI: https://doi.org/10.1007/s10278-020-00334-0
76. **Liu S, Wang Y, Wen A, Wang L, Hong N, Shen F,** et al. Implementation of a cohort retrieval system for clinical data repositories using the observational medical outcomes partnership common data model: proof-of-concept system validation. *JMIR Med Inform.* 2020; 8(10): e17376. DOI: https://doi.org/10.2196/17376
77. **Moore SM, Musil CM, Alder ML, Pignatiello G, Higgins P, Webel A,** et al. Building a research data repository for chronic condition self-management using harmonized data. *Nurs Res.* 2020; 69(4): 254–63. DOI: https://doi.org/10.1097/NNR.0000000000000435
78. **O'Neil ME, Harik JM, McDonagh MS, Cheney TP, Hsu FC, Cameron DC,** et al. Development of the PTSD-repository: a publicly available repository of randomized controlled trials for posttraumatic stress disorder. *J Trauma Stress.* 2020; 33(4): 410–9. DOI: https://doi.org/10.1002/jts.22520

79. **Ozaydin B, Zengul F, Oner N, Feldman SS.** Healthcare research and analytics data infrastructure solution: a data warehouse for health services research. *J Med Internet Res.* 2020; 22(6): e18579. DOI: https://doi.org/10.2196/18579
80. **Reimer AP, Milinovich A.** Using UMLS for electronic health data standardization and database design. *J Am Med Inform Assoc.* 2020; 27(10): 1520–8. DOI: https://doi.org/10.1093/jamia/ocaa176
81. **Samra H, Li A, Soh B.** GENE2D: A NoSQL integrated data repository of genetic disorders data. *Healthcare (Basel).* 2020; 8(3). DOI: https://doi.org/10.3390/healthcare8030257
82. **Uddin MA, Stranieri A, Gondal I, Balasubramanian V.** Rapid health data repository allocation using predictive machine learning. *Health Informatics J.* 2020; 26(4): 3009–36. DOI: https://doi.org/10.1177/1460458220957486
83. **Yu YW, Weber GM.** Balancing accuracy and privacy in federated queries of clinical data repositories: algorithm development and validation. *J Med Internet Res.* 2020; 22(11): e18735. DOI: https://doi.org/10.2196/18735
84. **Zubke M, Katzensteiner M, Bott OJ.** Integration of unstructured data into a clinical data warehouse for kidney transplant screening – challenges & solutions. *Stud Health Technol Inform.* 2020; 270: 272–6.
85. **Das S, Abou-Haidar R, Rabalais H, Sun S, Rosli Z, Chatpar K,** et al. The C-BIG repository: an institution-level open science platform. *Neuroinformatics*; 2021. DOI: https://doi.org/10.1007/s12021-021-09516-9
86. **Epstein RH, Hofer IS, Salari V, Gabel E.** Successful implementation of a perioperative data warehouse using another hospital's published specification from Epic's electronic health record system. *Anesth Analg.* 2021; 132(2): 465–74. DOI: https://doi.org/10.1213/ANE.0000000000004806
87. **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: https://doi.org/10.1186/s13054-021-03733-z
88. **Lenert LA, Ilatovskiy AV, Agnew J, Rudsill P, Jacobs J, Weatherston D,** et al. Automated production of research data marts from a canonical Fast Healthcare Interoperability Resource (FHIR) data repository: applications to COVID-19 research. *medRxiv.*; 2021. DOI: https://doi.org/10.1101/2021.03.11.21253384
89. **Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ,** et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. *J Am Med Inform Assoc.* 2021; 28(7): 1591–9. DOI: https://doi.org/10.1093/jamia/ocaa340
90. **Mahdi A, Błaszczyk P, Dłotko P, Salvi D, Chan TS, Harvey J,** et al. OxCOVID19 database, a multimodal data repository for better understanding the global impact of COVID-19. *Sci Rep.* 2021; 11(1): 9237. DOI: https://doi.org/10.1038/s41598-021-88481-4
91. **Majeed RW, Fischer P, Günther A.** Accessing OMOP common data model repositories with the i2b2 webclient – algorithm for automatic query translation. *Stud Health Technol Inform.* 2021; 278: 251–9. DOI: https://doi.org/10.3233/SHTI210077
92. **Majeed RW, Stöhr MR, Günther A.** HIStream-Import: a generic ETL framework for processing arbitrary patient data collections or hospital information systems into HL7 FHIR bundles. *Stud Health Technol Inform.* 2021; 278: 75–9. DOI: https://doi.org/10.3233/SHTI210053
93. **Schaaf J, Chalmers J, Omran H, Pennekamp P, Sitbon O, Wagner TOF,** et al. The registry data warehouse in the European reference network for rare respiratory diseases – background, conception and implementation. *Stud Health Technol Inform.* 2021; 278: 41–8. DOI: https://doi.org/10.3233/SHTI210049
94. **Shahid A, Nguyen TN, Kechadi MT.** Big data warehouse for healthcare-sensitive data applications. *Sensors (Basel).* 2021; 21(7). DOI: https://doi.org/10.3390/s21072353
95. **Guerriero L, Ferdeghini EM, Viola SR, Porro I, Testi A, Bedini R.** Telematic integration of health data: a practicable contribution. *Inform Health Soc Care.* 2011; 36(3): 147–60. DOI: https://doi.org/10.3109/17538157.2011.584997
96. **Bellazzi R, Masseroli M, Murphy S, Shabo A, Romano P.** Clinical bioinformatics: challenges and opportunities. *BMC Bioinformatics.* 2012; 13 Suppl 14(Suppl 14): S1. DOI: https://doi.org/10.1186/1471-2105-13-S14-S1
97. **Hulse NC, Galland J, Borsato EP.** Evolution in clinical knowledge management strategy at Intermountain Healthcare. *AMIA Annu Symp Proc.* 2012; 2012: 390–9.
98. **Taggart J, Liaw ST, Dennis S, Yu H, Rahimi A, Jalaludin B,** et al. The University of NSW electronic practice based research network: disease registers, data quality and utility. *Stud Health Technol Inform.* 2012; 178: 219–27.
99. **Huser V, Cimino JJ.** Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. *AMIA Annu Symp Proc.* 2013; 2013: 648–56.
100. **Jefferys BR, Nwankwo I, Neri E, Chang DC, Shamardin L, Hänold S,** et al. Navigating legal constraints in clinical data warehousing: a case study in personalized medicine. *Interface Focus.* 2013; 3(2): 20120088. DOI: https://doi.org/10.1098/rsfs.2012.0088
101. **Takecian PL, Oikawa MK, Braghetto KR, Rocha P, Lucena F, Kavounis K,** et al. Methodological guidelines for reducing the complexity of data warehouse development for transactional blood bank systems. *Decis Support Syst.* 2013; 55(3): 728–39. DOI: https://doi.org/10.1016/j.dss.2013.02.008

102. **Branescu I, Purcarea VL, Dobrescu R.** Solutions for medical databases optimal exploitation. *J Med Life.* 2014; 7(1): 109–18.
103. **Yoo S, Kim S, Lee KH, Jeong CW, Youn SW, Park KU,** et al. Electronically implemented clinical indicators based on a data warehouse in a tertiary hospital: its clinical benefit and effectiveness. *Int J Med Inform.* 2014; 83(7): 507–16. DOI: https://doi.org/10.1016/j.ijmedinf.2014.04.001
104. **Chosy J, Benson K, Belen D, Starr R, Lowery St John T, Starr RR,** et al. Insights in public health: for the love of data! The Hawai'i health data warehouse. *Hawaii J Med Public Health.* 2015; 74(11): 382–5.
105. **Cohen B, Vawdrey DK, Liu J, Caplan D, Furuya EY, Mis FW,** et al. Challenges associated with using large data sets for quality assessment and research in clinical settings. *Policy Polit Nurs Pract.* 2015; 16(3–4): 117–24. DOI: https://doi.org/10.1177/1527154415603358
106. **Pecoraro F, Luzi D, Ricci FL.** Data warehouse design from HL7 clinical document architecture schema. *Stud Health Technol Inform.* 2015; 213: 139–42.
107. **Pecoraro F, Luzi D, Ricci FL.** Designing ETL tools to feed a data warehouse based on electronic healthcare record infrastructure. *Stud Health Technol Inform.* 2015; 210: 929–33.
108. **Plazzotta F, Mayan JC, Storani FD, Ortiz JM, Lopez GE, Gimenez GM,** et al. Multimedia health records: user-centered design approach for a multimedia uploading service. *Stud Health Technol Inform.* 2015; 210: 474–8.
109. **Teixeira JW, Annibal LP, Felipe JC, Ciferri RR, Ciferri CD.** A similarity-based data warehousing environment for medical images. *Comput Biol Med.* 2015; 66: 190–208. DOI: https://doi.org/10.1016/j.compbiomed.2015.08.019
110. **Wyllie D, Davies J.** Role of data warehousing in healthcare epidemiology. *J Hosp Infect.* 2015; 89(4): 267–70. DOI: https://doi.org/10.1016/j.jhin.2015.01.005
111. **Aziz HA.** Handling big data in modern healthcare. *Lab Med.* 2016; 47(4): e38–e41. DOI: https://doi.org/10.1093/labmed/lmw038
112. **Haarbrandt B, Wilschko A, Marschollek M.** Modelling of operative report documents for data Integration into an openEHR-based enterprise data warehouse. *Stud Health Technol Inform.* 2016; 228: 407–11.
113. **Wanderer JP, Poler SM, Rothman BS.** Show me the data! A perioperative data warehouse of Epic Proportions. *Anesth Analg.* 2016; 122(6): 1742–3. DOI: https://doi.org/10.1213/ANE.0000000000001321
114. **Foran DJ, Chen W, Chu H, Sadimin E, Loh D, Riedlinger G,** et al. Roadmap to a comprehensive clinical data warehouse for precision medicine applications in oncology. *Cancer Inform.* 2017; 16: 1176935117694349. DOI: https://doi.org/10.1177/1176935117694349
115. **Jannot AS, Zapletal E, Avillach P, Mamzer MF, Burgun A, Degoulet P.** The Georges Pompidou University Hospital clinical data warehouse: A 8-years follow-up experience. *Int J Med Inform.* 2017; 102: 21–8. DOI: https://doi.org/10.1016/j.ijmedinf.2017.02.006
116. **Karami M, Rahimi A, Shahmirzadi AH.** Clinical data warehouse: an effective tool to create intelligence in disease management. *Health Care Manag (Frederick).* 2017; 36(4): 380–4. DOI: https://doi.org/10.1097/HCM.0000000000000113
117. **Kortüm KU, Müller M, Kern C, Babenko A, Mayer WJ, Kampik A,** et al. Using electronic health records to build an ophthalmologic data warehouse and visualize patients' data. *Am J Ophthalmol.* 2017; 178: 84–93. DOI: https://doi.org/10.1016/j.ajo.2017.03.026
118. **Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG,** et al. SMART-on-FHIR implemented over i2b2. *J Am Med Inform Assoc.* 2017; 24(2): 398–402. DOI: https://doi.org/10.1093/jamia/ocw079
119. **Dietrich G, Krebs J, Fette G, Ertl M, Kaspar M, Störk S,** et al. Ad hoc information extraction for clinical data warehouses. *Methods Inf Med.* 2018; 57(1): e22–e9. DOI: https://doi.org/10.3414/ME17-02-0010
120. **Fette G, Kaspar M, Liman L, Dietrich G, Ertl M, Krebs J,** et al. Exporting data from a clinical data warehouse. *Stud Health Technol Inform.* 2018; 248: 88–93.
121. **Mullin S, Zhao J, Sinha S, Lee R, Song B, Elkin PL.** Clinical data warehouse query and learning tool using a human-centered participatory design process. *Stud Health Technol Inform.* 2018; 251: 59–62.
122. **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: 94–9.
123. **Solbrig HR, Hong N, Murphy SN, Jiang G.** Automated population of an i2b2 clinical data warehouse using FHIR. *AMIA Annu Symp Proc.* 2018; 2018: 979–88.
124. **Tute E, Steiner J.** Modeling of ETL-processes and processed information in clinical data warehousing. *Stud Health Technol Inform.* 2018; 248: 204–11. DOI: https://doi.org/10.2196/13917
125. **Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ.** Building a semantic health data warehouse in the context of clinical trials: development and usability study. *JMIR Med Inform.* 2019; 7(4): e13917.
126. **Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML,** et al. eHOP Clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. *Stud Health Technol Inform.* 2019; 264: 1536–7.
127. **Oliveira PH, Scabora LC, Cazzolato MT, Oliveira WD, Paixao RS, Traina AJM,** et al. Employing domain indexes to efficiently query medical data from multiple repositories. *IEEE J Biomed Health*

Inform. 2019; 23(6): 2220–9. DOI: https://doi.org/10.1109/JBHI.2018.2881381

128. **Zohner J, Marquardt K, Schneider H, Michel Backofen A.** Challenges and opportunities in changing data structures of clinical document archives from HL7-V2 to FHIR-based archive solutions. *Stud Health Technol Inform.* 2019; 264: 492–5.
129. **Gavrilov G, Vlahu-Gjorgievska E, Trajkovik V.** Healthcare data warehouse system supporting cross-border interoperability. *Health Informatics J.* 2020; 26(2): 1321–32. DOI: https://doi.org/10.1177/1460458219876793
130. **Pavlenko E, Strech D, Langhof H.** Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med Inform Decis Mak.* 2020; 20(1): 157. DOI: https://doi.org/10.1186/s12911-020-01177-z
131. **Ronaldson A, Chandakas E, Kang Q, Brennan K, Akande A, Ebyarimpa I,** et al. Cohort profile: the East London Health and Care Partnership Data Repository: using novel integrated data to support commissioning and research. *BMJ Open.* 2020; 10(9): e037183. DOI: https://doi.org/10.1136/bmjopen-2020-037183
132. **Park J, You SC, Jeong E, Weng C, Park D, Roh J,** et al. A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. *JMIR Med Inform.* 2021; 9(3): e23983. DOI: https://doi.org/10.2196/23983
133. **Klann JG, McCoy AB, Wright A, Wattanasin N, Sittig DF, Murphy SN.** Health care transformation through collaboration on open-source informatics projects: integrating a medical applications platform, research data repository, and patient summarization. *Interact J Med Res.* 2013; 2(1): e11. DOI: https://doi.org/10.2196/ijmr.2454
134. **Rizi SA, Roudsari A.** Development of a public health reporting data warehouse: lessons learned. *Stud Health Technol Inform.* 2013; 192: 861–5.
135. **Cimino JJ, Ayres EJ, Remennik L, Rath S, Freedman R, Beri A,** et al. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS): design, contents, functionality and experience to date. *J Biomed Inform.* 2014; 52: 11–27. DOI: https://doi.org/10.1016/j.jbi.2013.11.004
136. **Marés J, Shamardin L, Weiler G, Anguita A, Sfakianakis S, Neri E,** et al. p-medicine: A medical informatics platform for integrated large scale heterogeneous patient data. *AMIA Annu Symp Proc.* 2014; 2014: 872–81.
137. **Pang X, Kozlowski N, Wu S, Jiang M, Huang Y, Mao P,** et al. Construction and management of ARDS/sepsis registry with REDCap. *J Thorac Dis.* 2014; 6(9): 1293–9.
138. **Schreiweis B, Schneider G, Eichner T, Bergh B, Heinze O.** Health Information Research Platform (HIReP)–an architecture pattern. *Stud Health Technol Inform.* 2014; 205: 773–7.
139. **Walji MF, Kalenderian E, Stark PC, White JM, Kookal KK, Phan D,** et al. BigMouth: a multi-institutional dental data repository. *J Am Med Inform Assoc.* 2014; 21(6): 1136–40. DOI: https://doi.org/10.1136/amiajnl-2013-002230
140. **Kunjan K, Toscos T, Turkcan A, Doebbeling BN.** A multidimensional data warehouse for community health centers. *AMIA Annu Symp Proc.* 2015; 2015: 1976–84.
141. **Narra L, Sahama T, Stapleton P.** Clinical data warehousing for evidence based decision making. *Stud Health Technol Inform.* 2015; 210: 329–33.
142. **Shenvi EC, Meeker D, Boxwala AA.** Understanding data requirements of retrospective studies. *Int J Med Inform.* 2015; 84(1): 76–84. DOI: https://doi.org/10.1016/j.ijmedinf.2014.10.004
143. **Obeid JS, Tarczy-Hornoch P, Harris PA, Barnett WK, Anderson NR, Embi PJ,** et al. Sustainability considerations for clinical and translational research informatics infrastructure. *J Clin Transl Sci.* 2018; 2(5): 267–75. DOI: https://doi.org/10.1017/cts.2018.332
144. **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): 1364–9. DOI: https://doi.org/10.1093/jamia/ocz068
145. **Artemova S, Madiot PE, Caporossi A, Mossuz P, Moreau-Gaudry A.** PREDIMED: clinical data warehouse of Grenoble Alpes University Hospital. *Stud Health Technol Inform.* 2019; 264: 1421–2.
146. **Chen YA, Tripathi LP, Fujiwara T, Kameyama T, Itoh MN, Mizuguchi K.** The TargetMine data warehouse: enhancement and updates. *Front Genet.* 2019; 10: 934. DOI: https://doi.org/10.3389/fgene.2019.00934
147. **Juárez D, Schmidt EE, Stahl-Toyota S, Ückert F, Lablans M.** A generic method and implementation to evaluate and improve data quality in distributed research networks. *Methods Inf Med.* 2019; 58(2–3): 86–93. DOI: https://doi.org/10.1055/s-0039-1693685
148. **Post A, Chappidi N, Gunda D, Deshpande N.** A method for EHR phenotype management in an i2b2 data warehouse. *AMIA Jt Summits Transl Sci Proc.* 2019; 2019: 92–101.
149. **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; 4(8): e17687. DOI: https://doi.org/10.2196/17687
150. **Williams M, Bagwell J, Zozus MN.** Data management plans: the missing perspective. *Journal*

of Biomedical Informatics. 2017; 71: 130–42. DOI: https://doi.org/10.1016/j.jbi.2017.05.004
151. **Keralis S, Stark S, Halbert M, Moen WE.** Research data management in policy and practice: the dataRes project. Research data management: principles, practice, and prospects. *Council on Library and Information Resources.* 2013; 160: 16–38.
152. **HDAA.** Available from: https://www.youtube.com/channel/UCLWx_z4hhfDUMv_XD_nermw.
153. **Kimball R, Ross M.** The data warehouse toolkit: the complete guide to dimensional modeling: John Wiley & Sons; 2011.
154. **Ross M, Kimball R.** The data warehouse toolkit: the definitive guide to dimensional modeling: John Wiley & Sons; 2013.
155. **Blaisure JC, Ceusters WM.** Improving the 'fitness for purpose' of common data models through realism based ontology. *AMIA Annu Symp Proc.* 2017; 2017: 440–7.
156. **McGinnis JM, Olsen L, Goolsby WA, Grossmann C.** Clinical data as the basic staple of health learning: creating and protecting a public good: workshop summary. National Academies Press; 2011.
157. **Gardner SP.** Ontologies and semantic data integration. *Drug Discovery Today.* 2005; 10(14): 1001–7. DOI: https://doi.org/10.1016/S1359-6446(05)03504-X
158. **Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU.** Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of Medical Informatics.* 2017; 26(1): 38–52. DOI: https://doi.org/10.15265/IY-2017-007
159. **Liyanage H, Krause P, Lusignan Sd.** Using ontologies to improve semantic interoperability in health data. *BMJ Health & Care Informatics.* 2015; 22(2): 309–15. DOI: https://doi.org/10.14236/jhi.v22i2.159
160. **Wang KC.** Standard lexicons, coding systems and ontologies for interoperability and semantic computation in imaging. *J Digit Imaging.* 2018; 31(3): 353–60. DOI: https://doi.org/10.1007/s10278-018-0069-8
161. **Takai-Igarashi T, Akasaka R, Suzuki K, Furukawa T, Yoshida M, Inoue K,** et al. On experiences of i2b2 (Informatics for integrating biology and the bedside) database with Japanese clinical patients' data. *Bioinformation.* 2011; 6(2): 86–90. DOI: https://doi.org/10.6026/97320630006086
162. **Cimino JJ.** Review paper: coding systems in health care. *Methods Inf Med.* 1996; 35(4–5): 273–84. DOI: https://doi.org/10.1055/s-0038-1634682
163. **Sanders D, Burton DA, Protti D.** The healthcare analytics adoption model: A framework and roadmap. *Health Catalyst.* 2013; 30.

How to cite this article: Wang Z, Syed M, Syed S, Greer M, Seker E, Zozus MN, Craven CK. Clinical Data Warehousing: A Scoping Review. *Journal of the Society for Clinical Data Management*. 2024; 4(1): 8, pp. 1–19. DOI: https://doi.org/10.47912/jscdm.320

Submitted: 14 November 2023 **Accepted:** 02 July 2024 **Published:** 25 July 2024

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS