

<https://doi.org/10.1038/s41746-025-01581-7>

A scoping review of OMOP CDM adoption for cancer research using real world data



Liwei Wang^{1,5}, Andrew Wen^{1,5}, Sunyang Fu¹, Xiaoyang Ruan¹, Ming Huang¹, Rui Li¹, Qiuhaio Lu¹, Heather Lyu², Andrew E. Williams^{3,4} & Hongfang Liu¹ ✉

The Observational Medical Outcomes Partnership (OMOP) common data model (CDM) supports large-scale research by enabling distributed network analyses. However, the breadth of its adoption in cancer research is not well understood. We conducted a scoping review to describe the adoption of the OMOP CDM in cancer research. A total of 49 unique articles were included in the review, with 30 on the data analysis theme, and 20 on the infrastructure theme. This review highlighted that while the OMOP CDM ecosystem has enabled successful data support for cancer research, particularly for collaborative studies, ongoing model development and iterative improvement remain needed to fulfill additional research data needs. Expanding disease sites, specifically for rare cancers, integrating more diverse types of data sources, improving data quality, adopting advanced analytics methodology, and increasing multisite evaluations serve as important opportunities to facilitate secondary usage of observational data in future cancer research.

Cancer has been a major cause of premature death globally¹ throughout the 21st century, prompting increasing research efforts and funding. Observational research has emerged as a powerful approach for generating hypotheses and uncovering unique insights into patient populations, treatments, and outcomes^{2,3}. This methodology has significantly advanced clinical understanding and influenced medical practices⁴. Primary sources of observational health data include electronic health records (EHRs), administrative claims, hospital billing systems, clinical registries, and longitudinal surveys⁵. Given the extensive role of observational health data, a robust framework for the generation of this information is essential to complete effective cancer studies and deliver high-quality cancer care.

Multicenter studies are widely utilized in observational research to improve the generalizability of findings. Distributed research networks, such as the Observational Health Data Sciences and Informatics (OHDSI)⁶, the Agency for Healthcare Research and Quality-supported projects⁷, the National Patient-Centered Clinical Research Network (PCORnet)⁸ and the Electronic Medical Records and Genomics (eMERGE) network, have emerged in recent years to promote multicenter observational studies⁹. Among these efforts, OHDSI supplies both a common data model (CDM) and the concept representation (terminology) for standardization to support federated analytics, demonstrating great potential for large-scale, collaborative observational cancer studies^{10,11}. The OHDSI network adopts the CDM developed as part of the Observational Medical Outcomes

Partnership (OMOP) to represent data from disparate sources in a standardized format through data normalization processes. The OHDSI network's CDM allows users to map data from disparate sources to a standardized format through data normalization processes. This enables a federated model, which allows individual data holders to maintain their patient-level databases locally, while allowing collaboration via systematic analytics, fostering diverse and geographically diverse patient cohorts, enhancing reproducibility, and ensuring patient confidentiality.

There have been two prior reviews on the role of OMOP CDM. One focused on the adoption of the OMOP CDM for observational patient data research between 2016 and 2021 and concluded that the relevance of the OMOP CDM is increasing regarding multi-country studies¹². Another review of the literature from 2016 through 2021 explored the potential application of the OMOP CDM in cancer prediction, specifically on the role of genomic vocabulary extensions in AI-based prediction models¹³. This study found that the OMOP CDM can enable a decentralized use of AI in the early prediction and diagnosis of cancer, personalized cancer treatment, and the discovery of important biological markers. While these studies have established the potential for the OMOP CDM, the breadth of its adoption in cancer research is not well understood. This paper aims to address this gap by presenting a scoping overview of the OMOP CDM in the field of cancer research to identify key opportunities and highlight unexplored areas for future investigation.

¹McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. ²Department of Surgical Oncology, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ³Clinical and Translational Science Institute, Tufts Medical Center, Boston, MA, USA. ⁴Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA. ⁵These authors contributed equally: Liwei Wang, Andrew Wen. ✉e-mail: hongfang.liu@uth.tmc.edu

Results

The complete article selection process is shown in Fig. 1. After identifying the included articles, the study team performed a comprehensive full-text review of the resulting 49 studies. There were 30 studies on data analysis and 20 studies on data infrastructure. Among them, one article (published in 2018) belonged to both themes¹⁴. All extracted data elements from the articles are provided in the Supplementary Data 1.

Overview analysis

The analysis of the included studies' metadata revealed insights into the distribution and trends of the research across the two themes: Infrastructure and Data Analysis. (Fig. 2) Although articles were collected from 2010 onward, the first article included in our study was published in 2017. There was an increasing trend in the publication of data analysis papers from 2018 to 2022 (Fig. 2a). Figure 2b compares the data sources used between Infrastructure and Data Analysis studies. One article may include more than one data source. EHR data were the most frequently used data source across both themes. Claims data served as another important source, particularly for the data analysis studies. EHR data were used in combination with another data source in six infrastructure-themed articles (claims and survey)^{15–20}, and seven data analysis-themed articles (claims, registry, and omics)^{10,21–26}. EHR data were used with two additional data sources (claims and registry) in only one infrastructure-themed article²⁷. Table 1 lists the references of data sources in each theme. EHR was the most commonly used data source in both the infrastructure and data analysis themes.

In terms of geographic distribution, North America, Asia, and Europe were the leading continents for article contribution, and a jump in accumulated publication numbers appeared in the year 2020 (Fig. 2c). Figure 2d illustrates a similar trend of cancer types studied across both themes, though there were more studies on blood and lymph cancers in the infrastructure theme. Table 2 provides detailed references of the specific cancer types and categorizations in each theme. Colorectal cancer was well studied in both themes.

Clusters based on cancer types and CDM table names were compared between the infrastructure and data analysis themes (Fig. 3). The infrastructure theme (Fig. 3a) covered all 8 CDM tables in the data analysis theme (Fig. 3b), and additionally incorporated several more tables, including Care_site, Cohort, Episode, Episode_event, Fact_relationship, Location, Note, Note_NLP, Device_exposure, and Specimen. It also covered extensions of the OMOP CDM including Genomic_test, Imaging_series, Imaging_study, Target_gene, Variant_annotation, Variant_occurrence, and

Vocabulary_extension. Such extensions of OMOP tables or vocabularies aimed to enable more comprehensive studies by addressing gaps that currently limit research due to the absence of necessary tables or vocabularies. For example, the Note_NLP table was included for colorectal cancer; Imaging_series and Imaging_study tables were included for prostate cancer; the Device_exposure table was included for breast and lung cancer; the Note table was included for thyroid cancer, and the Specimen table was included for blood, lung, and colorectal cancer. In the data analysis theme (Fig. 3b), the Condition_occurrence and Person tables were the most frequently used across all cancer types, followed by the Drug_exposure table, and the Observation, Measurement, and Procedure_occurrence tables. There was minimal use of the Visit_occurrence and Death tables.

Infrastructure theme

For all studies included in the infrastructure theme, efforts to develop reusable tooling and practices to transform cancer-specific data to the OMOP CDM format and to expand the OMOP CDM to support additional data were described. A total of 20 studies fell under this category. Studies in this category were divided into 4 subcategories (Table 3).

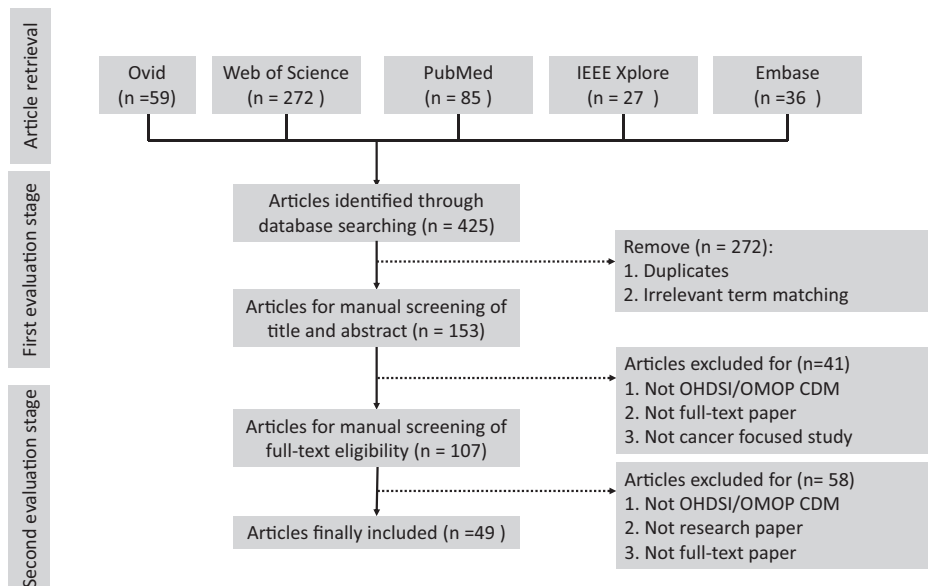
Studies in the infrastructure category were split relatively equally across three geographic regions including the United States ($n = 7$)^{16,17,20,27–30}, Europe ($n = 7$)^{15,19,31–35}, Asia with South Korea ($n = 5$)^{18,36–39} and China ($n = 1$)¹⁴. Within Europe, Germany was particularly distinct as it contributed five of the included studies from that region^{19,31–33,35}.

A majority of articles utilized a single dataset ($n = 11$)^{14,17,19,29–32,34,36–38}, which is reasonable for infrastructure construction efforts. One study reported four datasets²⁷, three studies involved three datasets^{16,33,39}, three studies involved two datasets^{18,20,28}, one study involved eight datasets³⁵, and one study involved 20 datasets¹⁵.

Of the studies ($n = 7$) that sought to extend the OMOP CDM or enrich the data contained within^{18,19,27,30,38,39}, four sought to extend the model to better support oncology-related data elements^{27,30,38,39}, two sought to extend support for –omics data^{18,19}, and one sought to extend support for imaging data³⁴.

A bulk ($n = 9$)^{14,16,17,20,32,34–37} of the studies in this category did not report a direct evaluation of the mapping quality into the OMOP CDM. Evaluation metrics were similarly ill-defined (Table 3 shows papers with some form of evaluation), although the most common evaluation was mapping coverage/percentage of source rows that were successfully mapped to the OMOP CDM standard ($n = 4$)^{28,29,38,39}, or the proportion of clinical concepts that could be successfully represented in the OMOP CDM standard ($n = 2$)^{27,31}. Two studies^{32,35} did not include an evaluation of the mapping process but did

Fig. 1 | Article selection process. Flow diagram illustrating the PRISMA approach for the identification, screening, and selection of studies.



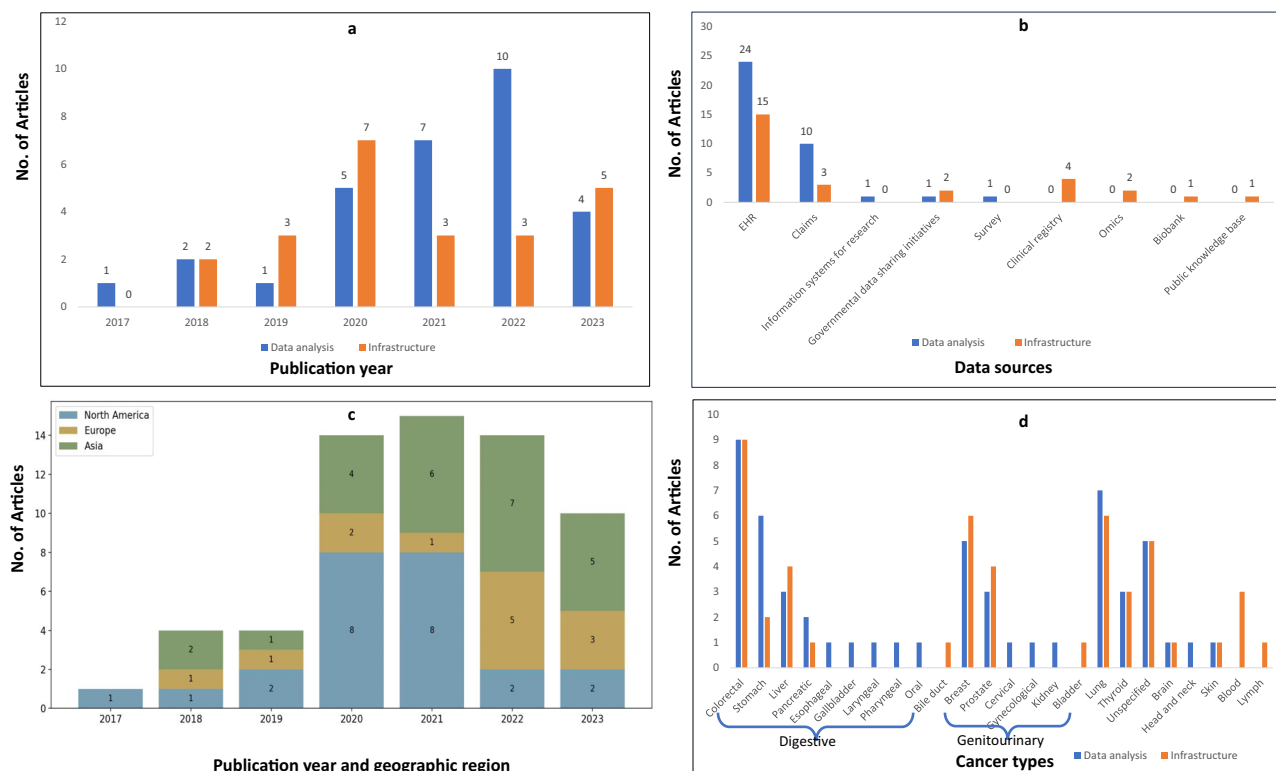


Fig. 2 | Distribution of all articles stratified by data analysis and infrastructure themes. Article distribution is shown across publication year (a), data sources (b), publication year and geographic region (c), and cancer types (d).

Table 1 | Comparison of Infrastructure and Data analysis in data sources

Data sources	Infrastructure (n = 20)	Data analysis (n = 30)
EHR	15 ^{14-20,27,28,33,34,36-39} (75%)	24 ^{10,11,14,21-26,40-49,55-57,59,60} (80%)
Claims	3 ^{16,25,27} (15%)	10 ^{10,21-23,25,26,51-54} (33.3%)
Information systems for research	0	1 ⁵⁰ (3.3%)
Governmental data sharing initiatives	2 ^{20,35} (10%)	1 ⁵⁸ (3.3%)
Survey	0	1 ²⁴ (3.3%)
Clinical registry	5 ^{15,17,27,31,32} (25%)	0
Omics	2 ^{18,19} (10%)	0
Biobank	1 ²⁹ (5%)	0
Public knowledge base	1 ³⁰ (5%)	0

report a metric of the percentage of concepts that were not represented in their tables.

Common themes regarding data mapping limitations were that the OMOP CDM could not represent certain clinically relevant concepts without further extension ($n = 6$)^{14,19,27,30,34,35} and some data were not directly available in structured form and required algorithmic normalization ($n = 3$)^{14,20,38}.

Data analysis theme

To better delineate the relationship amongst the various data elements collected, we conducted synthesis analyses for the data analysis theme. Figure 4 shows the linkage between aggregated cancer types, geographic area, study cohort size, study start year, and the study period. To categorize

Table 2 | Comparison of Infrastructure and Data analysis in cancer types

Classification	Specific cancer type	Infrastructure (n = 20)	Data analysis (n = 30)
Digestive system	Colorectal	6 ^{16,17,33,35,37,38} (30%)	9 ^{25,40,41,45,48,51,54,55,60} (30%)
	Stomach	2 ^{16,17} (10%)	6 ^{45,50-53,55} (20%)
	Liver	3 ^{14,16,17} (30%)	3 ^{45,51,55} (10%)
	Pancreatic	1 ¹⁷ (5%)	2 ^{45,51} (6.7%)
	Esophageal	0	1 ²¹ (3.3%)
	Gallbladder	0	1 ⁵¹ (3.3%)
	Oral-Laryn-Pharyngeal	0	1 ⁴⁴ (3.3%)
Genitourinary system	Breast	4 ^{16,28,32,33} (20%)	5 ^{25,45,51,55,59} (16.7%)
	Prostate	3 ^{16,34,36} (15%)	3 ^{45,49,51} (10%)
	Cervical	0	1 ²² (3.3%)
	Gynecological	0	1 ⁴⁵ (3.3%)
	Kidney	0	1 ⁵¹ (3.3%)
	Bladder	1 ²⁷ (5%)	0
Respiratory system	Lung	4 ^{16,18,32,36} (20%)	7 ^{24,25,42,45,51,55,57} (23.3%)
Endocrine system	Thyroid	3 ^{16,37,39} (15%)	3 ^{43,46,51} (10%)
Blood, blood vessel and lymph	Blood	3 ^{15,28,29} (15%)	0
	Blood vessel and Lymph	1 ³⁷ (5%)	0
Head and neck	Head and neck	0	2 ^{44,47} (6.7%)
Skin	Skin	1 ¹⁹ (5%)	1 ¹¹ (3.3%)
Unspecified	Unspecified	3 ^{20,30,31} (15%)	5 ^{10,23,26,56,58} (16.7%)

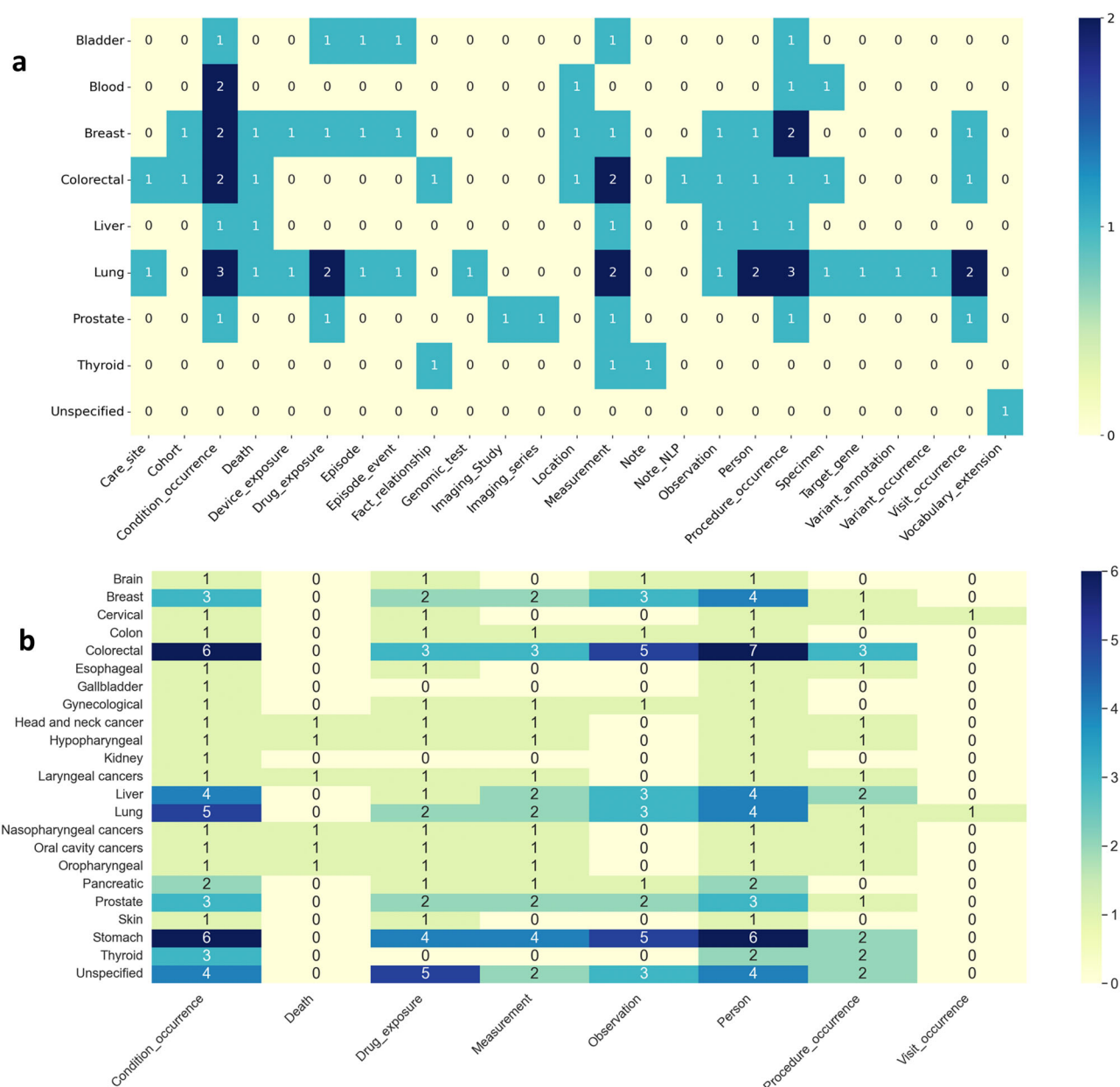


Fig. 3 | Comparison of clusters based on article numbers of co-occurrence of cancer types and CDM tables. a Infrastructure theme, **b** Data analysis theme. The infrastructure theme (**a**) covered all 8 CDM tables in the data analysis theme (**b**), and additionally incorporated several more tables, including Care_site, Cohort, Episode, Episode_event, Fact_relationship, Location, Note, Note_NLP, Device_exposure, and Specimen. It also covered extensions of the OMOP CDM including Genomic_test, Imaging_series, Imaging_study, Target_gene, Variant_annotation, Variant_occurrence, and Vocabulary_extension.

geographic locations, a global study is defined as a study that includes at least two countries, in contrast to a single-country study. Global studies ($n = 6$) began in 2020^{10,21,23,26,40,41}, and accounted for 20% of papers in the data analysis theme. Global collaborations were evident across multiple regions and countries, including USA, Spain, France, Germany, UK, Denmark, Netherlands, South Korea, and China, with the USA participating in the majority of studies, contributing to 5 out of 6 studies (83.3%). Among the 24 single-country studies, 15 came from South Korea, 6 from the USA, 2 from Denmark and 1 from China.

Among 30 studies in the data analysis theme, 15 (50%) studies leveraged multi-site datasets ranging from 2 to 11 individual sites^{10,11,21,23,25,26,41–49}. The remaining 15 studies used a single dataset, including 8 from South Korea^{50–57}, 4 studies from USA^{22,24,58,59}, and 1 each from Denmark⁶⁰, China¹⁴, and a collaboration effort between Denmark and Netherland⁴⁰. In terms of cancer types and population, 15 studies on

the South Korean population covered all cancer types except nervous system (brain cancer), which was exclusively conducted in the US population⁴⁷. Six local studies in the USA concentrated on genitourinary, nervous, and respiratory cancers^{22,24,44,47,49,59}. Denmark^{48,60} and China¹⁴ focused on digestive system cancers in their local studies. While global studies had the capacity to cover a population of more than one million patients^{23,26}, local studies included populations ranging from ≤ 1000 to 1 million. The earliest dataset started in 1986; two were from the South Korea^{25,44}, and one was an international, multisite study⁴⁹. The study periods of 8 studies exceeded 15 years^{10,23,25,44,46,48,49,60}. Four studies didn't provide the period of the studied population. Figure 5 shows the details of the distributions of cancer types by geolocation and study cohort size, top in-network institutions, and top out-network institutions compared with OHDSI collaborators (<https://www.ohdsi.org/who-we-are/collaborators/>).

Table 3 | A summary of papers in the infrastructure theme

Theme	Subcategory	Papers	Papers with some form of evaluation ^a
Infrastructure	Infrastructure development	$N = 7^{15-17,19,33,36,37}$	$N = 3^{15,19,33}$
	Transformation of various source data types to the OMOP CDM	$N = 5^{14,31,35,38,39}$	$N = 3^{31,38,39}$
	Data Model extensions and development	$N = 3^{18,27,34}$	$N = 2^{18,27}$
	Data Linkage and Standardization	$N = 5^{20,32}$	$N = 3^{28-30}$

^aData mapping evaluation metrics were ill-defined in the included papers, this column lists the papers with some form of evaluation.

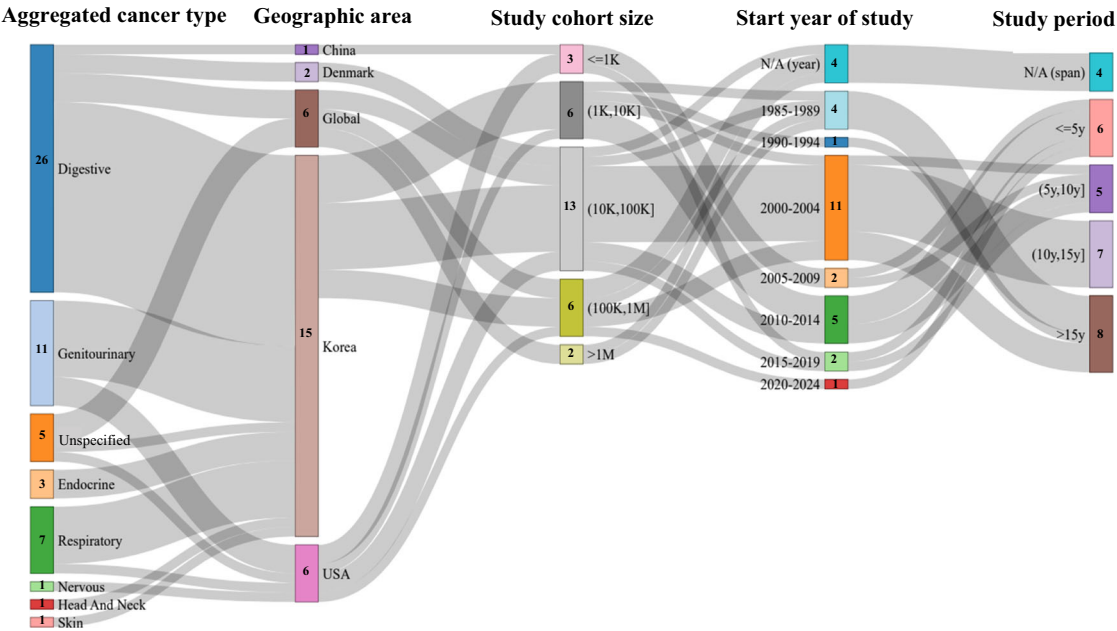


Fig. 4 | Linkage between the aggregated cancer type, geographic area, cohort size, start year of study, and study period. Numbers refer to paper numbers. A global study is defined as a study that includes at least two countries, in contrast to a single-country study.

Since studies from South Korea were disproportionately represented compared with other nations, Supplementary Fig. 1 shows the linkage after excluding studies from South Korea.

Study designs were categorized under two broader groups: “observational study” and “advanced analytics”. The “observational study” group was comprised of 22 (73.3%) papers, and “advanced analytics” group was comprised of 8 (26.7%) studies. Table 4 provides a list of the study methodologies.

Figure 6 illustrates the relationships between target domains, study designs, analysis methods, and CDM domain names used across all included data analysis papers. The majority (86.7%) of the research efforts focused on two primary target domains, i.e., diseases ($n = 15$)^{10,14,22,24,40,41,43,47-49,51,55,56,58,60} and drug-cancer association ($n = 11$)^{11,21,23,42,44-46,50,52-54,56}, respectively. Other domains included risk factors for emergency department (ED) visits⁵⁷, treatment patterns^{25,26}, and trial eligibility⁵⁹. Specifically, studies focusing on diseases are listed with their specific research questions in Table 5.

Other study designs in Fig. 6 include case control^{55,58}, cross sectional⁴³, and phenotyping⁵⁹. All 11 observational studies on drug-cancer association exclusively utilized the cohort study design. Conversely, observational studies on diseases included a variety of study designs. Among these, predictive modeling was the dominant approach ($n = 6$)^{40,41,47-49,60}, followed by cohort studies ($n = 4$)^{10,22,24,51}. The Cox regression model was the most widely used statistical method in observational studies ($n = 12$)^{11,14,21-23,42,44-46,50,52,55}, followed by logistic regression ($n = 5$)^{24,43,53,54,56}. Machine learning was the sole method for advanced

analytics in predictive modeling study design ($n = 7$). NLP was only employed for the trial eligibility via phenotyping⁵⁹. Supplementary Table 1 summarizes the studies using NLP. In data analysis papers, a wide range of CDM tables were analyzed by both statistical and machine learning methods, with the number of studies for each table shown in Fig. 6.

Discussion

In our review, cancer studies using OMOP CDM fell into two themes, data analyses and infrastructure construction. The presence of studies in both arenas indicates an ongoing evolution of OMOP integration into the data infrastructure for cancer researchers and centers. OHDSI was founded in 2008 and started to yield publications in 2010⁶¹, however, we found that studies with in-depth data analyses with OMOP data were not published until 2017⁴⁷, and publications in building out individual OMOP infrastructures were published in 2018. Global collaborative studies using OMOP started being published in 2020^{21,26,41}. Notably, the OMOP CDM enabled longitudinal studies with a study period spanning up to 15 years^{21,24,26,44,48,49,60} and projects with more than 1 million patients^{23,26}. Our review also demonstrated leaders in the field with the USA, South Korea, and Germany standing out as the leading countries leveraging the OMOP CDM for the cancer specific studies; this is consistent with a previous review⁶². Types of cancer research questions addressed in the data analytics studies varied widely and included disease-specific topics, drug-cancer association, risk factors for emergency department (ED) visits, treatment patterns, and clinical trial eligibility. Disease-specific and drug-cancer association were the most commonly studied topics. This demonstrates the

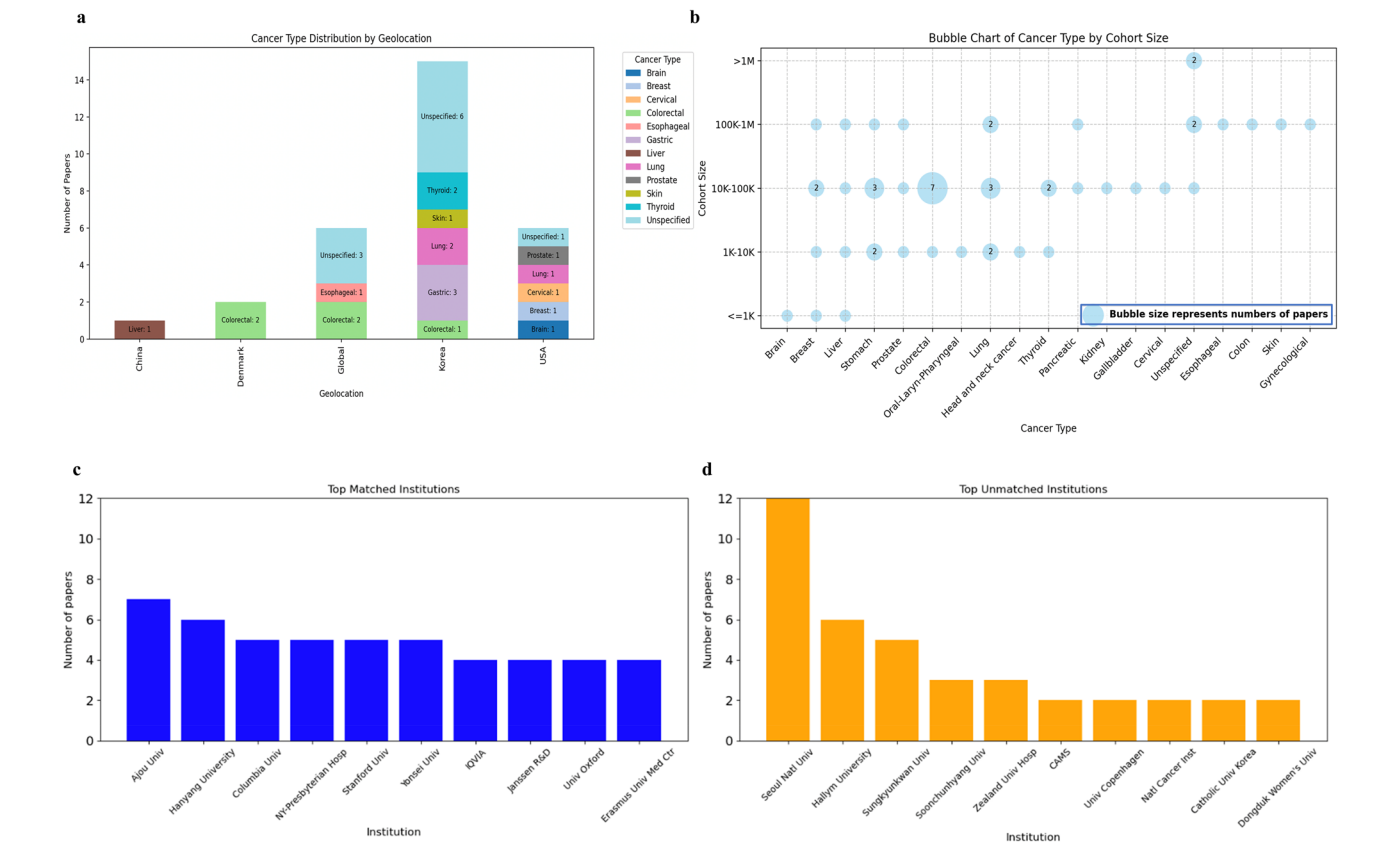


Fig. 5 | Distributions of cancer types. By geolocation (a), study cohort size (b), top in-network institutions (c), and top out-network institutions (d).

Table 4 | References for the study methods

Theme	Study methods (n = 30)	Subcategories	No. of studies with references
Data analysis	Observational study (n = 22, 73.3%)	Cohort study	n = 16 ^{10,11,21-24,42,44-46,50-54,56}
		Descriptive study	N = 3 ^{14,25,26}
		Case control	N = 2 ^{55,58}
		Cross sectional	N = 1 ⁴³
	Advanced analytics (n = 8, 26.7%)	Predictive modeling	N = 7 ^{40,41,47-49,57,60}
		Phenotyping	N = 1 ⁵⁹

potential utilization of the OMOP CDM for other types of cancer-specific topics such as drug repurposing and disease trajectory discovery.

To gain an understanding of whether real-world data are diverse enough and meet the data needs for downstream analysis, we investigated the cancer types in the studies across both themes. There was a wide range of cancer types covered in our review. However, when examining the cancer types, rare cancers were not well represented with limited studies on pancreatic cancer^{45,51} and pediatric brain cancer⁴⁷. The potential of OMOP CDM facilitating rare cancer science and discovery by pooling large-scale data is invaluable and warrants further exploration.

The diverse set of data sources included in the reviewed infrastructure studies suggests that cancer studies often require additional data sources including but not limited to clinical registries, omics, biobank, and population based datasets beyond the current EHR/Claims data-focused ecosystem. Meanwhile, new target CDM tables, such as Episode, Note_NLP and Specimen, and data model for omics and imaging data were extended in the infrastructure theme. It is evident that the OMOP CDM ecosystem is still undergoing active development and iteration, which will result in continuous improvements in its ability to support cancer research.

The reviewed studies in the data analysis theme were mostly observational cohort studies, demonstrating the important role of longitudinal analyses in generating hypotheses and showing important trends over time. Although limited in number, predictive studies using OMOP data were also highlighted in this review. Machine learning models were often used in these studies, while deep learning and large language model-based approaches remain yet unexplored. Advanced methodologies were also emphasized in the infrastructure theme – one study presented an overview of sustainable cloud-based platforms for developing, implementing, verifying, and validating trustable, usable, and reliable AI models for cancer care⁶³. Adopting advanced analytics methodologies will become important as data systems become more mature.

It should be noted that a substantial amount of clinically relevant information for cancer is represented in unstructured form. This is particularly true for certain types of data. For example, information within pathology reports is often difficult to capture, as synoptic reporting has been adopted for few cancer types at many institutions. However, limited studies explored the integration of NLP methods in building data infrastructure³⁷⁻³⁹, and only one study leveraged NLP-derived data in the data analysis theme⁵⁹. The potential challenges of current NLP methodologies for handling text data were highlighted in these studies, e.g., the limitations of using simple regex in NLP, along with concerns regarding generalizability and systematic evaluation of annotation schemas^{37,38,59}. We identified similar issues and barriers for wide adoption of cancer NLP in our previous study⁶⁴. Despite these challenges, it is critical to incorporate NLP-derived data within OMOP CDM instances for cancer research. A federated NLP deployment framework following the RITEFAIR (Reproducible, Implementable, Transparent, Explainable - Findable, Accessible, Interoperable, and Reusable) principles with scientific rigor and transparent (TRUST) provides a solution towards real-world clinical NLP while preserving the integrity and privacy for data from multiple sites^{65,66}.

Data quality challenges were typically attributed to two issues: accessibility information quality (IQ) and representational IQ^{67,68}. For

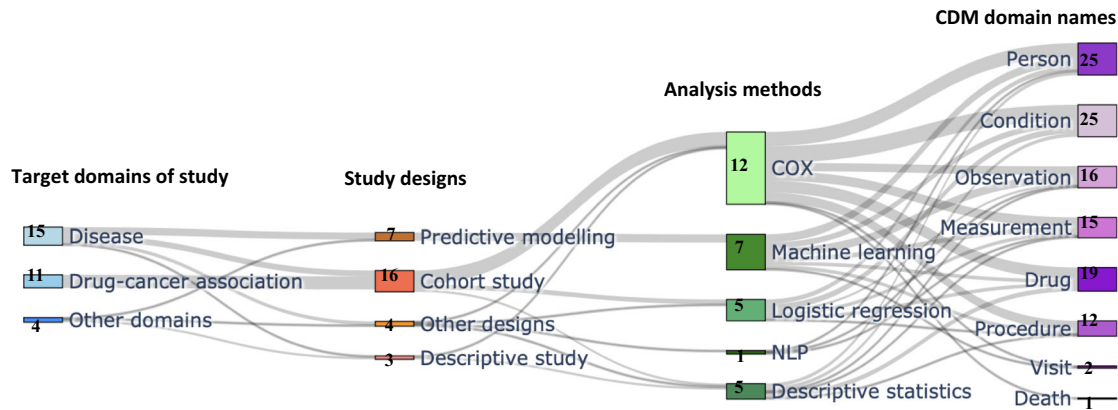


Fig. 6 | Analysis of target domains, study designs, analysis methods, and CDM domain names in the data analysis theme. Numbers refer to paper numbers. This figure illustrates the relationships between target domains, study designs, statistical methods, and CDM domain names used across all included data analysis papers.

Table 5 | Types of cancer research questions addressed in the disease target domain

Target domain	Research questions	Papers	
		Observational studies	Advanced analytics
Diseases	Disease characterization in cancer patients	<i>N</i> = 3 ^{10,24,51}	<i>N</i> = 0
	Cancer risk analysis	<i>N</i> = 1 ²²	<i>N</i> = 0
	Other disease risk analysis in cancer patients	<i>N</i> = 2 ^{56,58}	<i>N</i> = 0
	Cancer case prediction	<i>N</i> = 0	<i>N</i> = 2 ^{41,47}
	Perioperative disease analysis/prediction	<i>N</i> = 2 ^{43,55}	<i>N</i> = 2 ^{40,48}
	Survival analysis/prediction	<i>N</i> = 1 ¹⁴	<i>N</i> = 2 ^{41,60}

accessibility IQ, concerns related to poor record linkage and inaccessible geocoding information were discussed by several studies^{16,28,34}. Data timeliness was another issue as the current data retrieval and operation process is steward-based and lacks a real-time process (*n* = 2)^{17,33}. Data privacy, security (e.g., data identification), and regulatory considerations play a significant role in addressing accessibility IQ¹⁷. Regarding representational IQ, the lack of data standardization, particularly in the context of limited OMOP vocabularies, was noted as a challenge. In addition, a substantial portion of the reviewed studies in the infrastructure theme did not perform mapping quality evaluation. This is a significant issue as variations in this process can have profound effects on the validity of any downstream use cases. The potential solution for the data standardization and concept mapping problems lies in efforts to derive human-driven consensus amongst multiple use-cases on individual value-sets corresponding to individual clinical entities. Most prolific amongst these efforts is the NLM’s Value Set Authority Center (VSAC)⁶⁹ which aims to render clinical concept sets publicly available for further reuse and refinement. Beyond that, efforts have been made to create additional tooling allowing for similar functions at an institutional level (with greater human interaction), such as the OHNLP Valueset Workbench⁷⁰. Nevertheless, greater efforts should be made to integrate similar functionality into current clinical phenotyping workflows.

Although the OMOP CDM is designed to support multi-site studies, our review indicates that the majority of studies used single-site data. A gap in multisite evaluation for proposed methods/frameworks^{14,16,17,20,25,32,34–37,41,63} and representativeness of research findings due to single site data analysis design^{14,22,24,40,50–60} was observed in the infrastructure and data analysis themes, respectively. Site-specific biases within individual data sources further compound these challenges. Overall, the challenges lie in the multifaceted nature of the data ETL and harmonization processes, emphasizing the need for comprehensive and collaborative approaches to overcome technical, regulatory, and operational challenges.

While harmonization of clinical data via the OMOP CDM has vastly improved data standardization for multisite studies, these issues persist due to non-standard approaches by which these data are populated, particularly when it comes to concept normalization approaches. This issue is further complicated by the closed nature of many current EHR system licenses, limiting public sharing of developed ETL pipelines and leading to a substantial amount of re-implementation with differing methodologies. In the absence of any changes to EHR system licensing processes, the best approach is to actively publish concept mappings (e.g., via mechanisms such as the aforementioned Valueset Workbench⁶⁹) such that they can be reviewed, refined, and re-used by other collaborating institutions, particularly in the case of manual mappings and/or NLP-derived mappings from text-based clinical concepts⁷¹.

Limitations in our study included the potential biases of missing relevant articles that may be caused by search strings and databases selected, as well as the inherent ambiguity associated with data element collection, normalization, and analysis due to subjectivity occurred in the review process.

In conclusion, we conducted a scoping review to describe the adoption of the OMOP CDM in cancer research, providing an overview of efforts aimed at leveraging the OHDSI ecosystem for oncology studies. This review highlighted that while the OMOP CDM ecosystem has enabled successful data support for cancer research, particularly for collaborative studies, ongoing model development and iterative improvement remain needed to fulfill additional research data needs. Expanding disease sites, specifically for rare cancers, integrating more diverse types of data sources, improving data quality, adopting advanced analytics methodology, and increasing multisite evaluations serve as important opportunities to facilitate secondary usage of observational data in future cancer research.

Methods

We opted to employ a scoping review to explore the scope of the OMOP CDM for cancer research⁷². Our approach followed the framework outlined by Arksey and O’Malley⁷³, as well as the Preferred Reporting Items for

Systematic Reviews and Meta-Analyses (PRISMA) extension for scoping reviews⁷⁴. The process was conducted in five stages as detailed below:

Identifying the objectives

To analyze the status, challenges, and opportunities of adopting OMOP CDM for cancer research using real-world data, it's critical to have a complete understanding of the studies on practical applications of OMOP CDM, and those on data infrastructure construction. Therefore, we aimed to address the following objectives in the review process: (1) Examine the landscape of currently published cancer studies that utilized the OHDSI/OMOP CDM, (2) Assess the role of OHDSI/OMOP CDM as an appropriate data infrastructure for cancer research, and (3) Highlight challenges and opportunities to identify directions for future investigations.

Identifying relevant studies

We included articles published between January 1, 2010 through December 31, 2023. Articles written in English were retrieved from the following databases: Ovid, IEEE Xplore, PubMed, Web of Science, and Embase. A detailed description of the search strategies for articles using OHDSI OMOP for cancer related studies is provided in Supplementary Table 2.

Study selection

Two reviewers (L.W. and A.W.) independently screened the titles and abstracts of all articles retrieved. Publications were included if the OHDSI/OMOP CDM was used for cancer related studies. The following exclusion criteria were applied:

1. Non full-text papers
2. Articles retrieved by irrelevant term matching
3. Articles unrelated to OHDSI/OMOP CDM
4. Non-cancer articles
5. Non-research articles
6. Non-English language articles

A second round of full-text screening was performed by the same reviewers to ensure all publications met the inclusion and exclusion criteria. When disagreement arose, they discussed to achieve a consensus.

Charting the relevant studies

Four authors (L.W., A.W., S.F., H. Liu) designed the study themes, standardized templates for summarizing pertinent publications, and systematically organizing the information. Studies were categorized into two main themes: data analysis and infrastructure. The data analysis theme included observational studies or articles that utilized advanced analytics such as machine learning or natural language processing (NLP). In the infrastructure theme, we focused on studies describing reusable tools and practices for transforming data into the OMOP CDM format and expanding them to support additional data types, specifically in relation to cancer. Three reviewers (L.W., X.R., M.H.) were allocated for data element extraction for the data analysis theme, and three reviewers (A.W., Q.L., R.L.) were allocated for the infrastructure theme. Any disagreements were resolved by inter-discussion or discussion with a third reviewer to achieve a consensus.

The following described our review protocol. For the data analysis theme, data elements were extracted following the STROBE (strengthening the reporting of observational studies in epidemiology) checklist⁷⁵ a reporting guideline that describes core considerations for observational research. Key data elements included publication years, objectives, data sources, cancer type, institution names, geographic region, cohort size, target domain of study (disease, drug-cancer association, other.), method (machine learning, descriptive analysis, logistic regression, etc.), NLP usage (yes or no), study period (start year, end year), study design (cohort, case-control, and cross-sectional studies for observational study, predictive modeling, or phenotyping, etc.), variables analyzed (diagnosis, procedures, etc.), and number of datasets. To facilitate subsequent analyses, variables

were aggregated based on table names in OHDSI CDM version 5.4 (<https://ohdsi.github.io/CommonDataModel/>). Specifically, variables indicating lower-level clinical events were manually extracted from the method sections of included articles, and then manually mapped to the corresponding CDM tables. For example, medical history was mapped to the Condition_occurrence table, smoking status was mapped to the Observation table (see Supplementary Table 3). In addition, we extracted countries of the OMOP CDM datasets used for data analysis to identify geographic regions and institution names of the authors.

For the infrastructure theme, we categorized articles according to the OMOP CDM construction process (data linkage and standardization, transformation, etc.). Data elements of interest included publication year, geographic regions, institution names, cancer types, study topics, source data type (local EHR, claims data, etc.), target CDM table, mapping coverage, ETL (extract, transform, and load) challenges, mapping evaluation methods, data model extension, and limitations of data model (data element not specified, no definition, etc.). We extracted the countries of the authors to identify geographic regions and institution names of the datasets.

Collating, summarizing, and reporting the results

Data from the charting process were summarized, analyzed, and visualized to present an overview of the application of the OHDSI CDM in cancer research.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 20 November 2024; Accepted: 23 March 2025;

Published online: 07 April 2025

References

1. Bray, F., Laversanne, M., Weiderpass, E. & Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127**, 3029–3030 (2021).
2. Booth, C. M., Karim, S. & Mackillop, W. J. Real-world data: towards achieving the achievable in cancer care. *Nat. Rev. Clin. Oncol.* **16**, 312–325 (2019).
3. Baxter, N. N., Tepper, J. E., Durham, S. B., Rothenberger, D. A. & Virnig, B. A. Increased risk of rectal cancer after prostate radiation: a population-based study. *Gastroenterology* **128**, 819–824 (2005).
4. Callahan, A., Shah, N. H. & Chen, J. H. Research and reporting considerations for observational studies using electronic health record data. *Ann. Intern. Med.* **172**, S79–S84 (2020).
5. Voss, E. A. et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **22**, 553–564 (2015).
6. Hripcsak, G. et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud. Health Technol. Inform.* **216**, 574 (2015).
7. Randhawa, G. S. & Slutsky, J. R. Building sustainable multi-functional prospective electronic clinical data systems. *Med. Care* **50**, S3–S6 (2012).
8. Toh, S. et al. The National Patient-Centered Clinical Research Network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. *JMIR Res. Protoc.* **6**, e8323 (2017).
9. Gottesman, O. et al. The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
10. Roel, E. et al. Characteristics and outcomes of over 300,000 patients with COVID-19 and history of cancer in the United States and Spain. *Cancer Epidemiol. Biomark. Prev.* **30**, 1884–1894 (2021).
11. Lee, S. M. et al. Association between use of hydrochlorothiazide and nonmelanoma skin cancer: common data model cohort study in Asian population. *J. Clin. Med.* **9**, 2910 (2020).

12. Reinecke, I., Zoch, M., Reich, C., Sedlmayr, M. & Bathelt, F. The usage of OHDSI OMOP—a scoping review. *Ger. Med. Data Sci. 2021 Digital Med. Recognize—Understand—Heal* **21**, 95–103 (2021).
13. Ahmadi, N., Peng, Y., Wolfien, M., Zoch, M. & Sedlmayr, M. OMOP CDM can facilitate Data-driven studies for cancer prediction: a systematic review. *Int. J. Mol. Sci.* **23**, 11834 (2022).
14. Hong, N. et al. Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. *BMC Med. Inform. Decis. Mak.* **18**, 81–88 (2018).
15. Bardenheuer, K., Van Speybroeck, M., Hague, C., Nikai, E. & Price, M. Haematology Outcomes Network in Europe (HONEUR)—A collaborative, interdisciplinary platform to harness the potential of real-world data in hematology. *Eur. J. Haematol.* **109**, 138–145 (2022).
16. Cho, J. et al. Application of epidemiological geographic information system: an open-source spatial analysis tool based on the OMOP Common Data Model. *Int. J. Environ. Res. Public Health* **17**, 7824 (2020).
17. Glicksberg, B. S. et al. Blockchain-authenticated sharing of genomic and clinical outcomes data of patients with cancer: a prospective cohort study. *J. Med. Internet Res.* **22**, e16810 (2020).
18. Shin, S. J. et al. Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. *J. Med. Internet Res.* **21**, e13249 (2019).
19. Unberath, P. et al. EHR-independent predictive decision support architecture based on OMOP. *Appl. Clin. Inform.* **11**, 399–404 (2020).
20. Yu, Y. et al. Integrating electronic health record data into the ADEpedia-on-OHDSI platform for improved signal detection: a case study of immune-related adverse events. *AMIA Summits Transl. Sci. Proc.* **2020**, 710 (2020).
21. Kim, Y. et al. Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis. *Sci. Rep.* **10**, 11115 (2020).
22. Spotnitz, M. E., Natarajan, K., Ryan, P. B. & Westhoff, C. L. Relative risk of cervical neoplasms among copper and levonorgestrel-releasing intrauterine system users. *Obstet. Gynecol.* **135**, 319–327 (2020).
23. You, S. C. et al. Ranitidine use and incident cancer in a multinational cohort. *JAMA Netw. open* **6**, e2333495 (2023).
24. Na, J. et al. Characterizing phenotypic abnormalities associated with high-risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench. *J. Am. Med. Inform. Assoc.* **28**, 2313–2324 (2021).
25. Jeon, H. et al. Characterizing the anticancer treatment trajectory and pattern in patients receiving chemotherapy for cancer using harmonized observational databases: retrospective study. *JMIR Med. Inform.* **9**, e25035 (2021).
26. Chen, R. et al. Treatment patterns for chronic comorbid conditions in patients with cancer using a large-scale observational data network. *JCO Clin. Cancer Inform.* **4**, 171–183 (2020).
27. Belenkaya, R. et al. Extending the OMOP common data model and standardized vocabularies to support observational cancer research. *JCO Clin. Cancer Inform.* **5**, 12–20 (2021).
28. Jiang, X., Beaton, M. A., Gillberg, J., Williams, A. & Natarajan, K. Feasibility of linking areadeprivation index data to the OMOP common data model. In *AMIA Annual Symposium Proceedings*. **2022**, 587 (American Medical Informatics Association, 2023).
29. Michael, C. L., Sholle, E. T., Wulff, R. T., Roboz, G. J. & Champion, T. R. Jr Mapping local biospecimen records to the OMOP common data model. *AMIA Summits Transl. Sci. Proc.* **2020**, 422 (2020).
30. Warner, J. L. et al. HemOnc: a new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *J. Biomed. Inform.* **96**, 103239 (2019).
31. Carus, J., Nürnberg, S., Ückert, F., Schlüter, C. & Bartels, S. Mapping cancer registry data to the episode domain of the Observational Medical Outcomes Partnership Model (OMOP). *Appl. Sci.* **12**, 4010 (2022).
32. Carus, J. et al. Mapping the oncological basis dataset to the standardized vocabularies of a common data model: a feasibility study. *Cancers* **15**, 4059 (2023).
33. Gruendner, J. et al. KETOS: clinical decision support and machine learning as a service—A training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services. *PloS one* **14**, e0223010 (2019).
34. Kalokyri, V. et al. MI-Common Data Model: extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for registering medical imaging metadata and subsequent curation processes. *JCO Clin. Cancer Inform.* **7**, e2300101 (2023).
35. Maier, C. et al. Towards implementation of OMOP in a German university hospital consortium. *Appl. Clin. Inform.* **9**, 054–061 (2018).
36. Park, J., Lee, J. Y., Moon, M. H., Park, Y. H. & Rho, M. J. Cancer research line (CAREL): development of expanded distributed research networks for prostate cancer and lung cancer. *Technol. Cancer Res. Treat.* **22**, 15330338221149262 (2023).
37. Park, 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.* **9**, e23983 (2021).
38. Ryu, B. et al. Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. *J. Med. Internet Res.* **22**, e18526 (2020).
39. Yoo, S. et al. Transforming thyroid cancer diagnosis and staging information from unstructured reports to the observational medical outcome partnership common data model. *Appl. Clin. Inform.* **13**, 521–531 (2022).
40. Lin, V. et al. Training prediction models for individual risk assessment of postoperative complications after surgery for colorectal cancer. *Tech. Coloproctol.* **26**, 665–675 (2022).
41. Tian, Y. et al. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: a pilot study. *Int. J. Med. Inform.* **141**, 104173 (2020).
42. Lee, S. -H., et al. Angiotensin converting enzyme inhibitors and incidence of lung cancer in a population based cohort of common data model in Korea. *Sci. Rep.* **11**, 18576 (2021).
43. Lee, J. -H. et al. Assessment of inter-institutional post-operative hypoparathyroidism status using a common data model. *J. Clin. Med.* **10**, 4454 (2021).
44. Seol, S. et al. Effect of statin use on head and neck cancer prognosis in a multicenter study using a Common Data Model. *Sci. Rep.* **13**, 19770 (2023).
45. Lee, S. -H. et al. Renin-angiotensin-aldosterone system inhibitors and risk of Cancer: a population-based cohort study using a common data model. *Diagnostics* **12**, 263 (2022).
46. Kim, S. et al. Second primary malignancy risk in thyroid cancer and matched patients with and without radioiodine therapy analysis from the observational health data sciences and informatics. *Eur. J. Nucl. Med. Mol. Imaging* **49**, 3547–3556 (2022).
47. Felmeister, A. S. et al. Preliminary exploratory data analysis of simulated national clinical data research network for future use in annotation of a rare tumor biobanking initiative. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2098–2104 (IEEE, 2017).
48. Hartwig, M., Bräuner, K. B., Vogelsang, R. & Gögenur, I. Preoperative prediction of lymph node status in patients with colorectal cancer. Developing a predictive model using machine learning. *Int. J. Colorectal Dis.* **37**, 2517–2524 (2022).

49. Seneviratne, M. G., Banda, J. M., Brooks, J. D., Shah, N. H. & Hernandez-Boussard, T. M. Hernandez-Boussard. "Identifying cases of metastatic prostate cancer using machine learning on electronic health records. in *AMIA Annual Symposium Proceedings* **2018**, 1498 (American Medical Informatics Association, 2018).
50. Seo, S. I. et al. Association between proton pump inhibitor use and gastric cancer: a population-based cohort study using two different types of nationwide databases in Korea. *Gut* **70**, 2066–2075 (2021).
51. Yoon, J. Y., Kwak, M. S., Kim, H. I. & Cha, J. M. Seasonal variations in the diagnosis of the top 10 cancers in Korea: a nationwide population-based study using a common data model. *J. Gastroenterol. Hepatol.* **36**, 3371–3380 (2021).
52. Seo, S. I. et al. Aspirin, metformin, and statin use on the risk of gastric cancer: a nationwide population-based cohort study in Korea with systematic review and meta-analysis. *Cancer Med.* **11**, 1217–1231 (2022).
53. Kim, T. et al. Decreasing incidence of gastric cancer with increasing time after helicobacter pylori treatment: a nationwide population-based cohort study. *Antibiotics* **11**, 1052 (2022).
54. Seo, S. I. et al. Incidence and survival outcomes of colorectal cancer in long-term metformin users with diabetes: a population-based cohort study using a common data model. *J. Personalized Med.* **12**, 584 (2022).
55. Lee, Y. H., Kim, D. -H., Kim, J. & Lee, J. Risk assessment of postoperative pneumonia in cancer patients using a common data model. *Cancers* **14**, 5988 (2022).
56. Ha, H. et al. Application of the Khorana score for cancer-associated thrombosis prediction in patients of East Asian ethnicity undergoing ambulatory chemotherapy. *Thrombosis J.* **21**, 63 (2023).
57. Lee, A. R. et al. Risk prediction of emergency department visits in patients with lung cancer using machine learning: retrospective observational study. *JMIR Med. Inform.* **11**, e53058 (2023).
58. Song, Q. et al. Risk and outcome of breakthrough COVID-19 infections in vaccinated patients with cancer: real-world evidence from the National COVID Cohort Collaborative. *J. Clin. Oncol.* **40**, 1414 (2022).
59. Meystre, S. M., Heider, P. M., Kim, Y., Aruch, D. B. & Britten, C. D. Automatic trial eligibility surveillance based on unstructured clinical data. *Int. J. Med. Inform.* **129**, 13–19 (2019).
60. Bräuner, K. B. et al. Developing prediction models for short-term mortality after surgery for colorectal cancer using a Danish national quality assurance database. *Int. J. Colorectal Dis.* **37**, 1835–1843 (2022).
61. OHDSI. Observational Health Data Sciences and Informatics *OHDSI Publications*, <https://www.ohdsi.org/publications/>.
62. Bathelt, F. The usage of OHDSI OMOP—a scoping review. *Proce. German Med. Data Sci. (GMDS)*, 95–103 (2021).
63. Kondylakis, H. et al. Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. *Eur. Radiol. Exp.* **7**, 20 (2023).
64. Wang, L. et al. Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. *JCO Clin. Cancer Inform.* **6**, e2200006 (2022).
65. Liu, S. et al. An open natural language processing (NLP) framework for EHR-based clinical research: a case demonstration using the National COVID Cohort Collaborative (N3C). *J. Am. Med. Inform. Assoc.* **30**, 2036–2040 (2023).
66. Wen A. et al. The RECOVER Initiative. An NLP System for COVID/ PASC: A Case Demonstration of the OHNLP Toolkit from the National COVID Cohort Collaborative and the RECOVER programs. *JMIR Med. Inform.* **12**, e49997 (2024).
67. Lee, Y. W., Strong, D. M., Kahn, B. K. & Wang, R. Y. AIMQ: a methodology for information quality assessment. *Inf. Manag.* **40**, 133–146 (2002).
68. Fu, S. et al. The implication of latent information quality to the reproducibility of secondary use of electronic health records. *Stud. health Technol. Inform.* **290**, 173 (2022).
69. National Library of Medicine Value Set Authority Center. <https://vsac.nlm.nih.gov/>.
70. Peterson, K. J., Jiang, G., Brue, S. M., Shen, F. & Liu, H. Mining hierarchies and similarity clusters from value set repositories. in *AMIA Annual Symposium Proceedings*. **2017**, 1372 (American Medical Informatics Association, 2018).
71. Wen, A. et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. *NPJ Digital Med.* **2**, 130 (2019).
72. Munn, Z. et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* **18**, 1–7 (2018).
73. Arksey, H. & O'Malley, L. Scoping studies: towards a methodological framework. *Int. J. Soc. Res. Methodol.* **8**, 19–32 (2005).
74. Tricco, A. C. et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann. Intern. Med.* **169**, 467–473 (2018).
75. Von Elm, E. et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).

Acknowledgements

This project is supported by the Cancer Prevention Research Institute of Texas (CPRIT). RR230020, National Institute of Aging grant RF1AG072799, National Human Genome Research Institute R01HG12748, and National Library of Medicine R01LM11934. The funders played no role in the study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

L.W.: conceptualized and designed the study, analyzed the data of the data analysis theme, visualized results and drafted the manuscript; A.W.: conceptualized and designed the study, analyzed the data of the infrastructure theme and drafted the manuscript; F.S.: designed the study, analyzed the data of infrastructural theme and drafted the manuscript; X.R.: analyzed the data of the data analysis theme, visualized results and drafted the manuscript; M.H.: analyzed the data of the data analysis theme, visualized results and drafted the manuscript; R.L.: analyzed the data of infrastructural theme and visualized results; Q.L.: analyzed the data of infrastructural theme; A.E.W.: revised the manuscript; H.L.: revised the manuscript; H.L.*: conceptualized, supervised, designed the study and revised the manuscript.

Competing interests

Heather Lyu is an associate editor, and Hongfang Liu is an Editorial Board Member of NPJ Digital Medicine.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01581-7>.

Correspondence and requests for materials should be addressed to Hongfang Liu.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025