Title: Adoption of the OMOP CDM for Cancer Research using Real-world Data: Current Status and Opportunities

Authors:
Liwei Wang, MD, PhD[1]*, Andrew Wen, MS[1]*, Sunyang Fu, PhD[1], Xiaoyang Ruan, PhD[1], Ming Huang, PhD[1], Rui Li, PhD[1], Qiuhao Lu, PhD[1], Andrew E Williams, PhD[2,3], Hongfang Liu, PhD[1]

[1] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA
[2] Clinical and Translational Science Institute Tufts Medical Center Boston US
[3] Institute for Clinical Research and Health Policy Studies Tufts Medical Center Boston US

Liwei Wang and Andrew Wen contribute equally.

Corresponding author: Hongfang Liu, PhD
Postal address: 7000 Fannin Street #Suite 600, Houston, TX 77030
E-mail: hongfang.liu@uth.tmc.edu
Telephone: 713-500-3900

Keywords: Real-world data, cancer research, Observational Health Data Sciences and Informatics (OHDSI) network, Observational Medical Outcomes Partnership (OMOP), Common Data Model (CDM), scoping review

Word count: 3999

ABSTRACT

Background: The Observational Medical Outcomes Partnership (OMOP) common data model (CDM) that is developed and maintained by the Observational Health Data Sciences and Informatics (OHDSI) community supports large scale cancer research by enabling distributed network analysis. As the number of studies using the OMOP CDM for cancer research increases, there is a growing need for an overview of the scope of cancer research that relies on the OMOP CDM ecosystem.

Objectives: In this study, we present a comprehensive review of the adoption of the OMOP CDM for cancer research and offer some insights on opportunities in leveraging the OMOP CDM ecosystem for advancing cancer research.

Materials and Methods: Published literature databases were searched to retrieve OMOP CDM and cancer-related English language articles published between January 2010 and December 2023. A charting form was developed for two main themes, i.e., clinically focused data analysis studies and infrastructure development studies in the cancer domain.

Results: In total, 50 unique articles were included, with 30 for the data analysis theme and 23 for the infrastructure theme, with 3 articles belonging to both themes. The topics covered by the existing body of research was depicted.

Conclusion: Through depicting the status quo of research efforts to improve or leverage the potential of the OMOP CDM ecosystem for advancing cancer research, we identify challenges and opportunities surrounding data analysis and infrastructure including data quality, advanced analytics methodology adoption, in-depth phenotypic data inclusion through NLP, and multisite evaluation.

INTRODUCTION

Throughout the 21st century, cancer has been a major cause of premature death internationally[1], leading to substantial research interest. A promising avenue by which can be studied is via observational research, which holds great promise for generating real-world evidence and unique insights, e.g., into patients, treatments, and outcomes.[2,3] This avenue significantly contributes to advancing clinical knowledge and shaping medical practices.[4] The primary sources of observational health data encompass electronic health records (EHRs), insurance/administrative claims, hospital billing, clinical registries, and longitudinal surveys.[5] Given the promise shown in observational research, maximizing the potential of such data is crucial for effective cancer studies, high-quality cancer care, and improved cancer care management.

In particular, conducting multicenter studies is a common strategy used in observational clinical research that allows for improved generalizability of the results, and consequently, improved efficiency. To promote multicenter observational studies, distributed research networks have emerged in recent years, such as the Observational Health Data Sciences and Informatics (OHDSI),[6] the Agency for Healthcare Research and Quality (AHRQ)-supported projects,[7] the National Patient-Centered Clinical Research Network (PCORnet)[8] and the Electronic Medical Records and Genomics (eMERGE) network.[9] Among these efforts, OHDSI supplies both a common data model (CDM) and the concept representation (terminology) for standardization to support federated analytics, showing great potential for large-scale 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. A key benefit of such a network-based federated approach is that data holders can maintain their patient-level databases locally, allowing for collaboration through the distributed research network on systematic analytics, increased sample size, heterogeneous patient populations that are geographically dispersed and racially and ethnically diverse, enhanced research generalizability and reproducibility while still maintaining patient confidentiality.

Two previous related reviews have been done on the OMOP CDM. One focused on the adoption of the OMOP CDM in the field of observational patient data research, which delineated the trend over a 5-year period between 2016 and early 2021 by analyzing metadata and topics of literature.[12] Results confirmed the increasing importance of the OMOP CDM in conducting network studies internationally within the medical domain. Following that, another review investigated the potential applicability of the OMOP CDM in cancer prediction and how comprehensively the genomic vocabulary extension of the OMOP CDM can serve the needs of AI-based predictions based on the literature between 2016 and 2021.[13] This study found that the OMOP CDM serves as a solid base to enable a decentralized use of AI in early prediction, diagnosis, personalized cancer treatment, and in discovering important biological markers. While these studies have established the potential for the OMOP CDM for cancer research, the scope of the adoption of the OMOP CDM for cancer research is not well understood. This paper aims to bridge this gap by presenting a comprehensive outline for researchers in the field of cancer study leveraging the OMOP CDM, and guide them to several unexplored research gaps.

METHODS

Given our objective to explore the scope of the OMOP CDM for cancer studies, we opted for a scoping review. Scoping reviews have been described as an ideal tool for assessing the breadth and extent of a body of literature on a given topic, offering a comprehensive overview of its primary focus and coverage.[14] We conducted this scoping review with the following five stages based on the framework from Arksey and O'Malley,[15] and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for scoping reviews.[16]

Identifying the Research Question

In this scoping literature review, we aimed to identify 1) the extent of cancer data analysis utilizing the OHDSI/OMOP CDM, 2) the maturity of OHDSI/OMOP CDM as an ecosystem infrastructure for cancer research, and 3) challenges and opportunities from the above two themes for potential future investigations.

Identifying Relevant Studies

We included articles published from January 1, 2010 to December 31, 2023. Only studies written in English were considered. Literature databases surveyed included Journals@Ovid@TMC Library (subscribed full text), Journals@Ovid (some full text), Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations and Daily <1946 to January 12, 2024>; 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 ***Appendix 1***.

Study Selection

All the titles and abstracts after deduplication were screened, and the publications were included if OHDSI/OMOP CDM was used for cancer related studies. We excluded publications if they were

1. Not a full-text paper
2. Retrieved by irrelevant term matching
3. Not using OHDSI/OMOP CDM
4. Not cancer focused
5. Not a research paper
6. Not written in English

A second round of full-text screening was done to ensure all publications met the inclusion and exclusion criteria.

Charting the Relevant Studies

Standardized charting templates were created to summarize pertinent publications. The information of interest was organized around two main themes: data analysis and infrastructure. Two reviewers were assigned to each article, and tasked with independently extracting the information. Consensus was achieved after discussing disagreements between the two reviewers or consultation with a third reviewer.

Shared data elements extracted from the two themes include publication year, data sources, geographic region, and cancer type.

The data analysis theme includes both observational studies and data mining studies. We developed our data extraction schema partially based on the STROBE (strengthening the reporting of observational studies in epidemiology) checklist[17], a reporting guideline that describes core considerations for observational research. Data elements to extract include objectives, geographic region, cohort size, target domain (disease, drug, etc.), analysis type (SQL, machine learning, statistical analysis, etc.), OHDSI tool used, study period, study design (cohort, case-control, and cross-sectional studies for observational study, ML, or phenotyping, etc.), risk factors explored if applicable, variables (diagnosis, procedures, etc. based on the OHDSI CDM table names), statistical methods, NLP usage, number of datasets. To facilitate subsequent analysis, we aggregated variables based on the OHDSI CDM table names (https://ohdsi.github.io/CommonDataModel/).

A data extraction schema for the infrastructure theme was developed to encompass key components including source data warehouse type (local EHR, claims data, etc.), source data type (diagnoses, procedures, etc. based on the OHDSI CDM table names), mapping coverage, main challenges in ETL, evaluation method of mapping, data model extension, limitation of data model (data element not specified, no definition, etc.) and entity linking/normalization method.

Collating, Summarizing, and Reporting the Results

The results obtained from the data charting for each theme were summarized, analyzed, and visualized to present an overview of the application of OHDSI/CDM in the field of cancer.

RESULTS

The article selection process is shown in Figure 1. After identifying the included articles, the study team performed a comprehensive full-text review of the resulting 50 studies. There are 30 studies focusing on data analysis and 23 on data infrastructure. Among them, 3 articles (published in 2018, 2020, and 2021) belong to both data analysis and infrastructure. [18-20] All extracted data from the articles (charting items) are provided in the ***Data Supplement***.

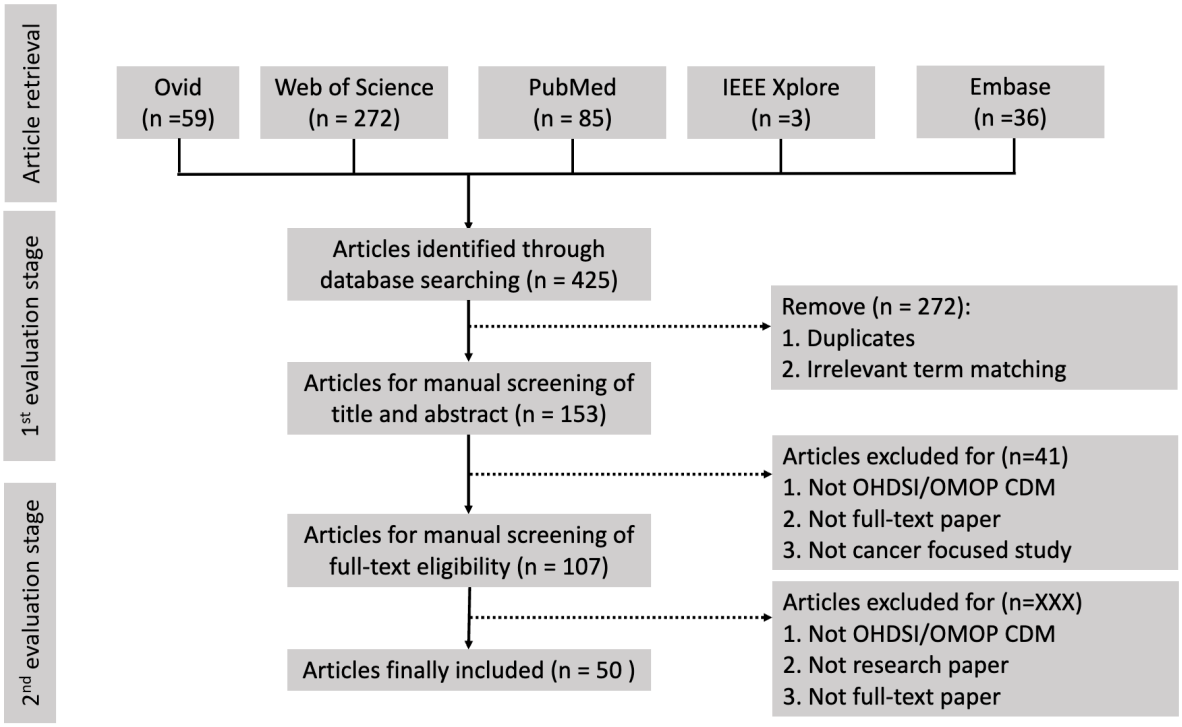


Figure 1. Article selection process.

Overview analysis

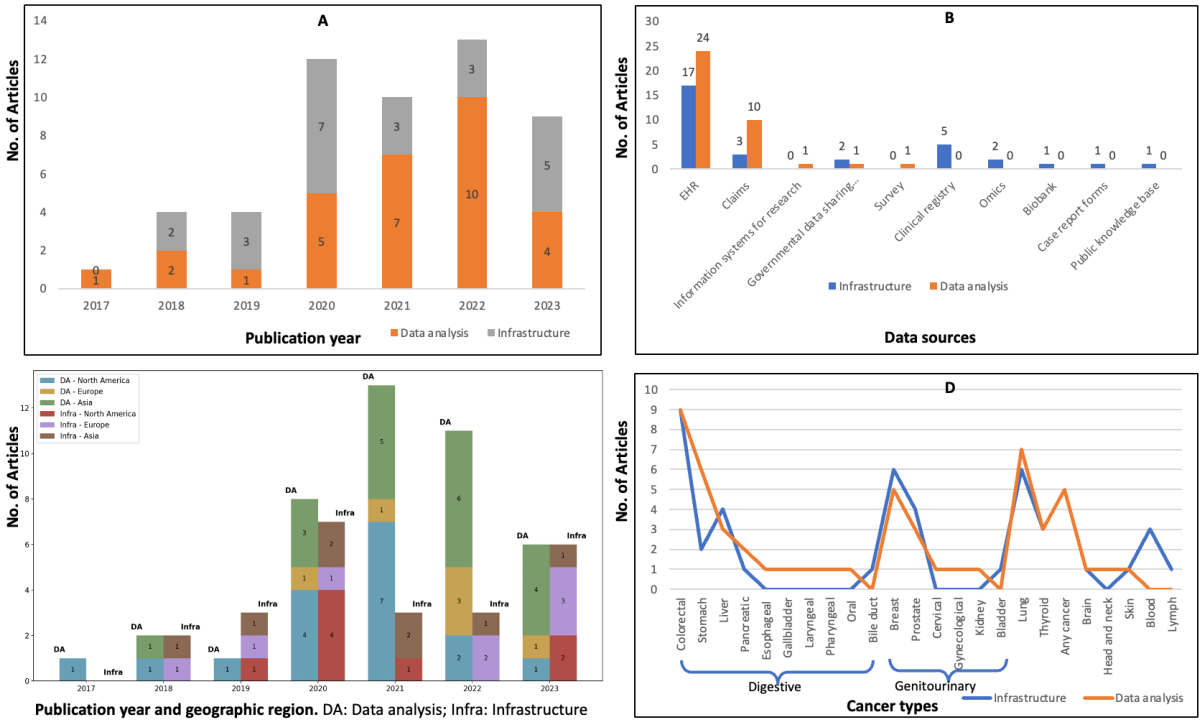


Figure 2. A: Distribution of all articles across publication year (A), data sources (B), publication year and geographic region (C) and cancer types (D), stratified by data analysis and infrastructure.

Figure 2 shows the distribution of studies stratified by the two themes, i.e., Infrastructure and Data analysis, with 3 articles (published in 2018, 2020, 2021) belonging to both data analysis and infrastructure. Though we collected articles from 2010, the first article included in our study was published in 2017, and data analysis papers showed an increasing trend from 2018 to 2022 (Figure 2A).

Figure 2B compares the data sources used between Infrastructure and Data Analysis studies. One article may include more than one data source. In general, usage of EHR data has been the mainstream in both themes, with claims data being another important source for the data analysis theme. EHR was used in combination with one additional data source in 7 infrastructure-themed articles (claims and survey),[19,21-26] and 7 data analysis-themed articles (claims, registry, and omics).[10,19,27-31] EHR was used with two additional data sources (claims and registry) in only 1 infrastructure-themed article. [32] Compared with the data analysis theme, several new types of data sources emerged for infrastructure construction, including clinical registries, omics, Biobank, case report forms, and public knowledge bases. Table 1 lists the references of data sources in each theme.

Table 1: Comparison of Infrastructure and Data analysis in data sources.

Data sources	Infrastructure (n=23)	Data analysis (n=30)
EHR	17 [18-26,32-39] (74%)	24 [10,11,18-20,27-31,40-53] (80%)
Claims	3 [19,22,32] (13%)	10 [10,19,27-29,31,54-57] (33.3%)
Information systems for research	0	1 [58] (3.3%)
Governmental data sharing	2 [24,59] (8.7%)	1 [60] (3.3%)
Survey	0	1 [30] (3.3%)
Clinical registry	5 [21,25,32,61,62] (21.7%)	0
Omics	2 [23,26] (8.7%)	0
Biobank	1 [63] (4.3%)	0
Case report forms	1 [64] (4.3%)	0
Public knowledge base	1[65] (4.3%)	0

Figure 2C shows the distribution of papers geographically in the North America, Asia and Europe. The USA, South Korea, and Germany stood out as the leading countries in each geographic region in the infrastructure and data analysis themes. More detailed analyses are shown in the results of the infrastructure and data analysis themes below. Figure 2D shows a similar trend of cancer types between the data analysis and the infrastructure theme, with a spike in the infrastructure theme highlighting the need for data construction for blood and lymph. The cancer types studied in the two themes covered a broad range, and the variation in the number of articles focused on each type was present. Table 2 lists the references (n>1) of the specific cancer types in each theme.

Table 2: Comparison of Infrastructure and Data analysis in cancer types.

Classification	Specific cancer	Infrastructure	Data analysis
Digestive system	Colorectal	9 [18,19,22,25,37-39,59,64]	9 [18,19,43,44,50-52,54,57]
	Stomach	2 [22,25]	6 [43,44,54-56,58]
	Liver	4 [22,25,64,66]	3 [43,44,54]
	Pancreatic	1 [25]	2 [43,54]
Genitourinary system	Breast	6 [19,22,34,39,62,64]	5 [19,43,44,48,54]
	Prostate	4 [22,33,35,64]	3 [43,53,54]
Respiratory system	Lung	6 [19,22,26,35,62,64]	7 [19,30,40,43,44,49,54]
Endocrine system	Thyroid	3 [22,36,37]	3 [41,45,54]
Blood and blood forming organs	Blood	3 [21,34,63]	0

Figure 3 compared the clusters based on cancer types and CDM table names (variables) between the infrastructure and Data analysis themes. Compared with the data analysis theme (Figure 3B), richer variable tables were involved in the infrastructure theme (Figure 3A) for colorectal cancer including Episode, Episode_event, Fact_relationship, Location, Note, Note_NLP, Specimen, Care_site, Unique_conditions, Uniqe_observations and Unique_procedures; for breast cancer including Device_exposure, lymphovascular cancer including vascular Note, thyroid cancer including Note, and blood cancer including specimen.

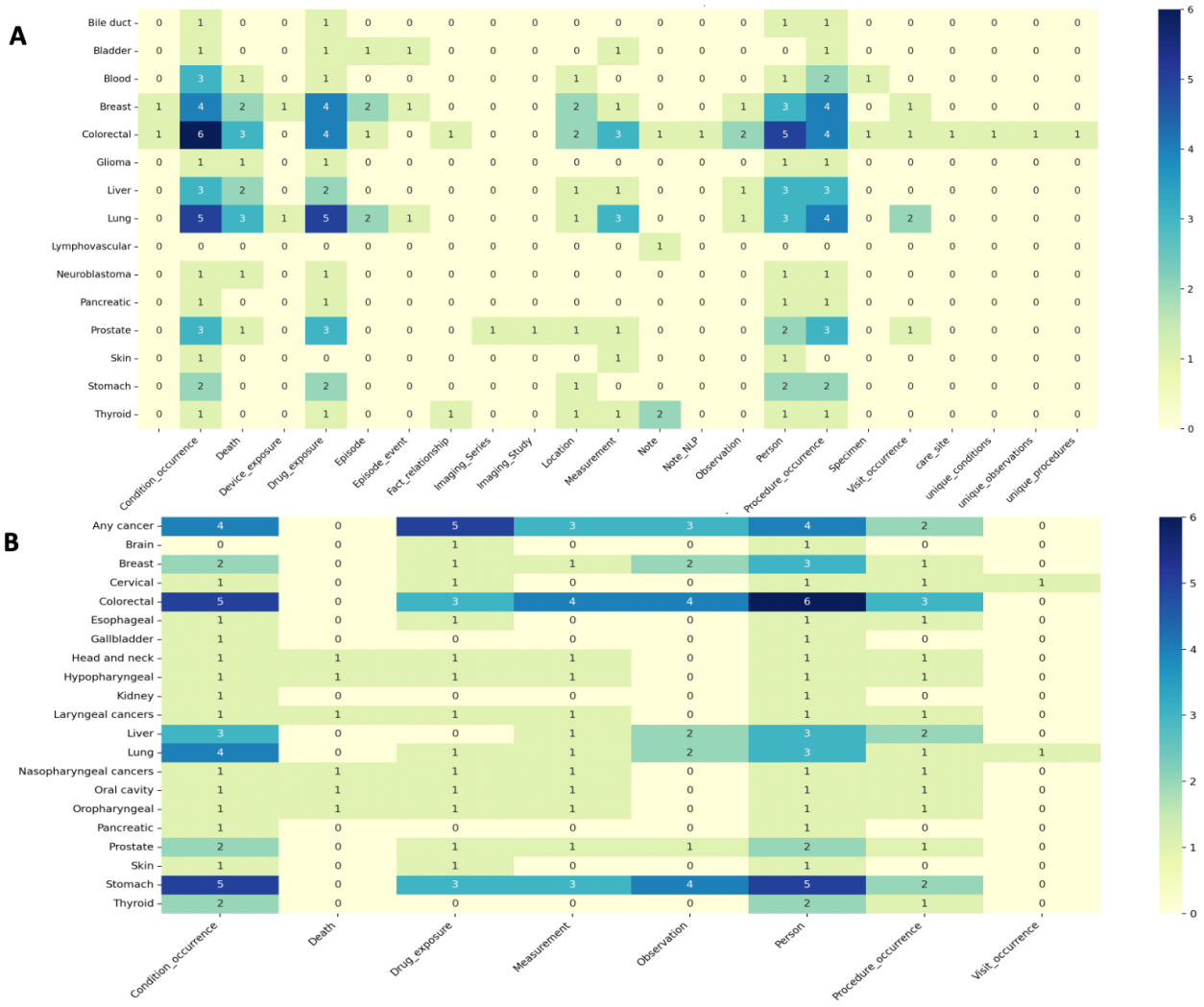


Figure 3. Comparison of clusters based on article numbers of co-occurrence of cancer types and CDM tables (variables). A: Infrastructure theme, B: Data analysis theme.

Infrastructure theme

Before data analysis can be conducted, the data itself must be present in the OMOP CDM format, and tooling to support that data analysis must exist. In this theme, we will therefore summarize efforts to develop reusable tooling and practices to transform data to the OMOP CDM format, as well as expand the OMOP CDM to support additional data, in relation to cancer. A total of 23 studies fell under this category. Broadly speaking, studies done in this category can be divided into 4 subcategories, i.e., infrastructure development, transformation of various source data types to the OMOP CDM, Data Model extensions and development, and Data Linkage and Standardization. Table 3 shows the references of papers in the 4 subcategories.

Table 3. A summary of papers in the infrastructure theme

Theme	Subcategory	Papers	Papers with some form of evaluation

Infrastructure	infrastructure development	N=9 [18,21-23,25,35,37,39,64]	N=3 [21,23,39]
	transformation of various source data types to the OMOP CDM	N=6 [19,20,36,38,59,61]	N=3 [36,38,61]
	Data Model extensions and development	N=3 [26,32,33]	N=2 [26,32]
	Data Linkage and Standardization	N=5 [24,34,62,63,65]	N=3 [34,63,65]

Geographic region and datasets

In terms of geographic region, studies within this category are split equally across three geographic regions with the United States (n=8), [22,24,25,32,34,63-65] Europe (n=8), [21,23,33,39,59,61,62,64] Asia with South Korea (n=6) [19,26,35-38] and China (n=2).[18,66] Within Europe, Germany is particularly distinct as it participates in 5 of the included studies from that region. [23,39,59,61,62]

A majority of these articles remain concentrated within a single dataset (n=11), [23,25,33,35,37,38,61-63,65,66] which is reasonable for infrastructure construction efforts. Of the remainder, 4 studies involve 3 datasets, [19,22,36,39] 3 studies involve 2 datasets, [24,26,34] 1 studies involve 6, [18] 1 study involves 8, [32] and 1 study involves 20. [21] One study did not report a dataset. [33]

OMOP data and model extension

Of the studies (n=8) that sought to extend the OMOP CDM or enrich the data contained within, [23,26,32,33,36,38,64,65] 5 sought to extend the model to better support oncology-related data elements, [32,36,38,64,65] 2 sought to extend support for –omics data, [23,26] and 2 sought to extend support for imaging data. [33,64]

Data mapping and evaluation

A bulk (n=12) [18,19,22,24,25,33,35,37,59,62,64,66] of the studies in this category do not report a direct evaluation of the mapping quality into the OMOP CDM. Evaluation metrics were similarly ill-defined, although the most common evaluation was mapping coverage/percentage of source rows that were successfully mapped to the OMOP CDM standard (n=4), [34,36,38,63] or the proportion of clinical concepts that could be successfully represented in the OMOP CDM standard (n=2). [32,61] Besides the studies reporting evaluation (n=11), two studies [59,62] did not have an evaluation of the mapping process but did report a metric of the percentage of concepts that were not representable.

Common themes regarding reported limitations of data mapping include the fact that the OMOP CDM could not represent certain clinically relevant concepts without further extension (n=6) [23,32,33,59,65,66] and that some data was not directly available in structured form and required algorithmic normalization (n=3). [24,38,66]

Data analysis theme

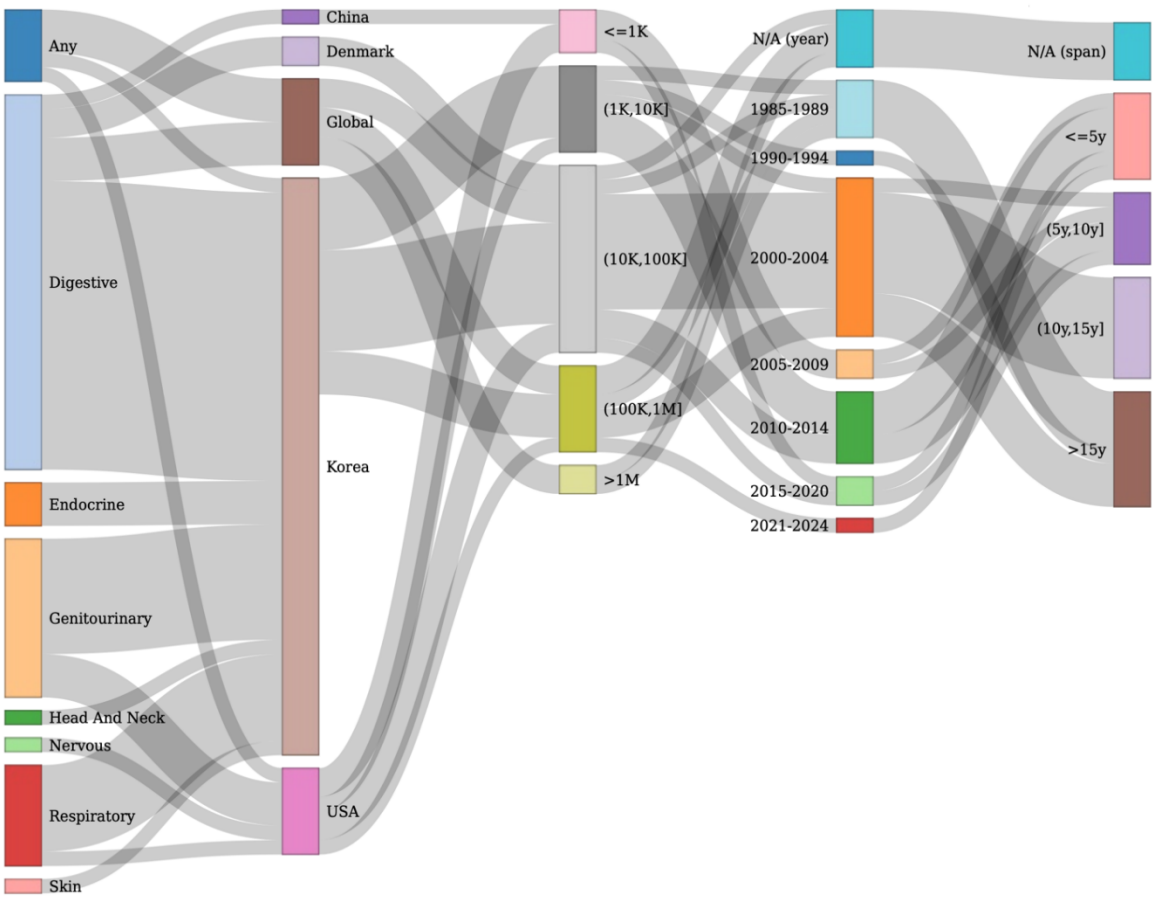


Figure 4. Linkage between the aggregated cancer type, geographic area, cohort size, start year of study, and study period. Analysis based on all countries.

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 population size, and the study period of the corresponding population. To categorize geographic locations, the global study is defined as a study that includes at least two countries, in contrast to a single-country study, which includes only one country. Global studies (n=6) started from 2020, [10,18,27,29,31,52] that accounted for 20% of papers in the data analysis theme. Global collaborations were across North America, Europe, and Asia, 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. [10,11,18,19,27,29,31,40-43,45,47,51,53] The remaining 15 studies used a single dataset, including 8 from South Korea [44,46,49,54-58] 4 studies from USA, [28,30,48,60] and 1 each from Denmark, [50] China, [20] and a collaboration effort between Denmark and Netherland. [52] 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 population. [47] Six local studies in the USA concentrated on genitourinary, nervous, and respiratory

cancers. [28,30,42,47,48,53] Denmark [50,51] and China [20] focused on digestive system cancers in their local studies. While global studies had the capacity to cover more than 1M population, [29,31] local studies covered the population ranging from <=1K to 1M. The earliest period started in 1986; two covered the South Korean population, [19,42] and one covered the global population. [49] Study period of 7 studies exceeded 15 years. [27,30,31,42,50,51,53] Four studies didn't provide the period of the studied population.

As studies from South Korea are disproportionally prevalent compared with other nations, to simplify visualization, Appendix Figure 2 shows the linkage after excluding local studies from South Korea. The upper right of the figure indicates that the study period between 1995 and 1999, and three cancer types, i.e., endocrine, head and neck and skin cancers were dropped out compared with Figure 3A.

Study designs are categorized under two broader groups: "observational study" and "advanced analytics". The "observational study", comprising 22 (73.3%) papers, and "advanced analytics" was presented in the relatively minor portion with 8 (26.7%) studies. Table 4 provides a list of references for the study methods.

Table 4. References for the study methods.

Theme	Study methods (n=30)	Subcategories	Papers with some form of evaluation
Data analysis	Observational study (n=22, 73.3%)	Cohort study	n=16 [10,11,27-30,40,42,43,45,46,54-58]
		Descriptive study	N=3 [19,20,31]
		Case control	N=2 [44,60]
		Cross sectional	N=1 [41]
	Advanced analytics (n=8, 26.7%)	Predictive modeling	N=7 [18,47,49-53]
		Phenotyping	N=1 [48]

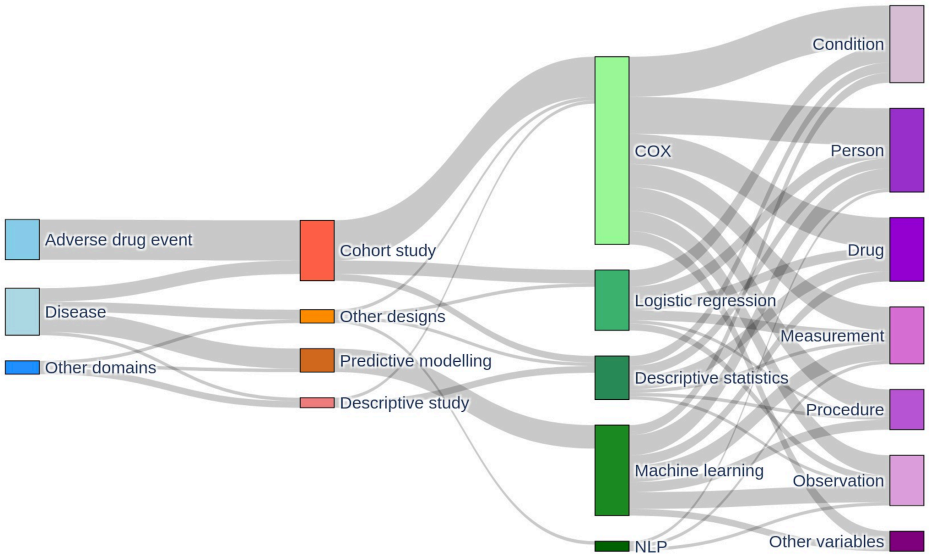


Figure 5. Analysis of target domains, study designs, statistical methods, and variables in the data analysis theme

Figure 5 illustrates the relationships between target domains, study designs, statistical methods, and variables used across these data analysis studies. The majority (86.7%) of the research efforts focused on two primary domains, i.e., disease (n=14) [10,18,20,28,30,41,44,47,50-54,60] and adverse drug events (ADE) (n=12), [11,27,29,40,42,43,45,46,55-58] respectively. Other domains included risk factors for emergency department (ED) visits, [49] treatment patterns, [19,31] and trial eligibility [48].

All the 14 observational studies on ADE exclusively utilized the cohort study design. Conversely, observational studies on diseases include a variety of study designs. Among these, predictive modeling was the dominant approach (n=6), [18,47,50-53] while cohort studies ranked second in usage (n=4). [10,28,30,54] Specifically, the COX model was the most widely used statistical method in observational studies (n=12), [11,20,27-29,40,42-45,55,58] followed by logistic regression (n=5). [30,41,46,56,57] Machine learning is the sole method for advanced analytics in predictive modeling study design (n=7). NLP was only employed in an observational study for the trial eligibility via phenotyping. [48]

In data analysis studies, a wide range of variables were leveraged by both statistical and machine learning methods. The most frequently used variables include condition occurrence, person, drug exposure, death, procedure occurrence, measurement, observation, and visit occurrence.

DISCUSSION

We conducted a scoping review on the adoption of the OMOP CDM for cancer studies since 2010. In the following subsections, we will discuss the extent of cancer data analysis and the maturity of the OMOP CDM as an infrastructural eco-system for cancer research, as well as associated challenges and opportunities for future investigation.

Status quo

The existence of data analysis-themed studies implies that the data used was prepared sufficiently for the targeted studies while infrastructure-themed studies might imply unmet data management needs. OHDSI was founded in 2008 and started to yield publications in 2010, [67] while cancer data analysis studies started in 2017, [47] and infrastructure publications started in 2018, global studies started in 2020 [18,27,31]. Of note, OMOP CDM enabled longitudinal studies spanning 15 years of study period [27,30,31,42,50,51,53] and studies with more than 1 million population. [29,31] It's shown that the USA, South Korea, and Germany stood out as the leading countries improving or leveraging OMOP CDM for the cancer domain in each continent, consistent with the previous review study. [12] It's worth noting that most data analytics studies focused on disease and adverse drug events.

Maturity of the OMOP CDM ecosystem

Examining the cancer types, data sources, and variable CDM tables being studied is helpful in understanding whether real-world data are well-prepared and meet the data needs for downstream analysis. Infrastructure and data analysis showed a roughly consistent trend in the wide range of cancer types they covered. The diverse set of data sources included in the reviewed infrastructure studies suggests that cancer studies require additional data sources beyond the current EHR data-focused ecosystem. Meanwhile, new variable tables, such as Episode, Note_NLP and Specimen, and data model extension for omics and imaging data were involved in the infrastructure theme. It is, therefore, evident that the OMOP CDM ecosystem is still undergoing active development and iteration, which will result in continuous improvement to better support cancer studies.

Adoption of advanced analytics methodology

Cancer research of the data analysis theme showed a strong preference for observational cohort studies, placing high value on long-term longitudinal analysis for drawing evidence over time. While limited in number, data mining studies were explored to gain predictive insights, suggesting an emerging stream in cancer research within the OMOP CDM framework. Machine learning models were the primary methods, while deep learning and large language model-based approaches remain unexplored. In light of the critical role of data infrastructure, one study presented an overview of the development efforts towards sustainable AI cloud-based platforms for developing, implementing, verifying, and validating trustable, usable, and reliable AI models regarding cancer care provision. [64]

In-depth phenotypic data inclusion

It should be noted that a substantial amount of clinically relevant information for cancer is represented in unstructured form. This is particularly the case for information contained within pathology reports, as synoptic reporting is only currently adopted for a minority of cancer types within many institutions. However, limited studies explored NLP methods to build data infrastructure, [36-38] and only 1 study leveraged NLP-derived data in the data analysis theme. [48] Potential challenges of the current NLP methodology 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. [48,37,38] We also identified and discussed issues and barriers for wide adoption of cancer NLP in our previous study. [68] Despite the challenges, it is critical to incorporate NLP-derived data within OMOP CDM instances for cancer research. In the context of multiple sites and privacy-preserving demands, a federated NLP

deployment framework following the RITE-FAIR (Reproducible, Implementable, Transparent, Explainable - Findable, Accessible, Interoperable, and Reusable) principles with scientific rigor and transparent (TRUST) provides a solution towards real-world clinical NLP. [69,70]

Data quality

The data quality challenge was primarily related to two sub-types, i.e., accessibility information quality (IQ) and representational IQ. [71,72] For accessibility IQ, concerns related to poor record linkage and inaccessible geocoding information were discussed by several studies. [22,33,34] Data timeliness was another issue as the current data retrieval and operation process is steward-based and lacks a real-time process (n=2). [25,39] Data privacy, security (e.g., data reidentification) and regulatory considerations play a significant role in addressing accessibility IQ. [25] Regarding representational IQ, the lack of data standardization, particularly in the context of limitations within OMOP vocabularies, was a continuous challenge. In addition, a substantial portion of the reviewed studies in the infrastructure theme did not perform mapping quality evaluation, which presented a significant gap as variations in such a process can have profound effects on the validity of any downstream use cases. The potential solution for the data standardization and concept mapping problem 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) [73] 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 Workbench. [74,75] Nevertheless, each of these tools is relatively standalone and greater effort should be made to integrate similar functionality into current clinical phenotyping workflows.

Multisite evaluation

Despite the OMOP CDM is designed to support multi-site studies, our review indicates that the majority of studies used single-site data. A lack of multisite evaluation for proposed methods/frameworks,[18,19,22,24,25,33,35,37,59,62,64,66] and representativeness of research findings due to single site data analysis design. [28,30,44,46,48,49,54-58,60,50, 52,20] was shown in the infrastructure and data analysis themes, respectively. Site-specific infrastructural biases within individual data sources further compound these challenges. Overall, the challenges lie in the multifaceted nature of the data ETL and harmonization process, emphasizing the need for comprehensive approaches to overcome technical, regulatory, and operational challenges.

While harmonization of clinical data via the OMOP CDM has vastly improved this state of affairs, the issue persists due to non-standard approaches by which this data is 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 change on the EHR license terms, perhaps the best approach is to actively publish concept mappings (e.g., via mechanisms such as the aforementioned Valueset Workbench [73]) such that they can be reviewed, refined, and re-used later on down the line, particularly in the case of manual mappings and/or NLP-derived mappings from text-based clinical concepts.

CONCLUSION

In this scoping review, we depicted the status quo of research efforts to improve or leverage the potential of the OMOP CDM ecosystem for advancing cancer research. Our findings revealed that while the OMOP CDM ecosystem has reached a level of maturity that is sufficient to support cancer research, ongoing model development and iteration remains needed to fulfill additional research data needs. Subsequently, we identify challenges and opportunities surrounding data analysis and infrastructure including data quality, advanced analytics methodology adoption, in-depth phenotypic data inclusion through NLP, and multisite evaluation.

ACKNOWLEDGEMENT

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.

AUTHOR CONTRIBUTION

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.: designed the study, 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. Wi.: conceptualized, and revised the manuscript; H.L.: conceptualized, supervised, and designed the study and revised the manuscript.

REFERECES

1. Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer *127*, 3029-3030.
2. Booth, C.M., Karim, S., and Mackillop, W.J. (2019). Real-world data: towards achieving the achievable in cancer care. Nature reviews Clinical oncology *16*, 312-325.
3. Baxter, N.N., Tepper, J.E., Durham, S.B., Rothenberger, D.A., and Virnig, B.A. (2005). Increased risk of rectal cancer after prostate radiation: a population-based study. Gastroenterology *128*, 819-824.
4. Callahan, A., Shah, N.H., and Chen, J.H. (2020). Research and reporting considerations for observational studies using electronic health record data. Annals of internal medicine *172*, S79-S84.
5. Voss, E.A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, F.J., Londhe, A., Zhu, V., and Ryan, P.B. (2015). Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association *22*, 553-564.

6. Hripcsak, G., Duke, J.D., Shah, N.H., Reich, C.G., Huser, V., Schuemie, M.J., Suchard, M.A., Park, R.W., Wong, I.C.K., and Rijnbeek, P.R. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics *216*, 574.
7. Randhawa, G.S., and Slutsky, J.R. (2012). Building sustainable multi-functional prospective electronic clinical data systems. Medical Care, S3-S6.
8. Toh, S., Rasmussen-Torvik, L.J., Harmata, E.E., Pardee, R., Saizan, R., Malanga, E., Sturtevant, J.L., Horgan, C.E., Anau, J., and Janning, C.D. (2017). The National Patient-Centered Clinical Research Network (PCORnet) bariatric study cohort: rationale, methods, and baseline characteristics. JMIR research protocols *6*, e8323.
9. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., and Basford, M.A. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. Genetics in Medicine *15*, 761-771.
10. Roel, E., Pistillo, A., Recalde, M., Sena, A.G., Fernández-Bertolín, S., Aragón, M., Puente, D., Ahmed, W.-U.-R., Alghoul, H., and Alser, O. (2021). Characteristics and outcomes of over 300,000 patients with COVID-19 and history of cancer in the United States and Spain. Cancer Epidemiology, Biomarkers & Prevention *30*, 1884-1894.
11. Lee, S.M., Kim, K., Yoon, J., Park, S.K., Moon, S., Lee, S.E., Oh, J., Yoo, S., Kim, K.-I., and Yoon, H.-J. (2020). Association between use of hydrochlorothiazide and nonmelanoma skin cancer: Common data model cohort study in Asian population. Journal of Clinical Medicine *9*, 2910.
12. Bathelt, F. (2021). The usage of OHDSI OMOP–a scoping review. Proceedings of the German Medical Data Sciences (GMDS), 95.
13. Ahmadi, N., Peng, Y., Wolfien, M., Zoch, M., and Sedlmayr, M. (2022). OMOP CDM can facilitate Data-Driven studies for cancer prediction: A systematic review. International Journal of Molecular Sciences *23*, 11834.
14. Munn, Z., Peters, M.D., Stern, C., Tufanaru, C., McArthur, A., and Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology *18*, 1-7.
15. Arksey, H., and O'Malley, L. (2005). Scoping studies: towards a methodological framework. International journal of social research methodology *8*, 19-32.
16. Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D., Horsley, T., and Weeks, L. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine *169*, 467-473.
17. Von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C., and Vandenbroucke, J.P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet *370*, 1453-1457.
18. Tian, Y., Chen, W., Zhou, T., Li, J., Ding, K., and Li, J. (2020). Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: A pilot study. International Journal of Medical Informatics *141*, 104173.
19. Jeon, H., You, S.C., Kang, S.Y., Seo, S.I., Warner, J.L., Belenkaya, R., and Park, R.W. (2021). Characterizing the anticancer treatment trajectory and pattern in patients

receiving chemotherapy for cancer using harmonized observational databases: Retrospective study. JMIR Medical Informatics *9*, e25035.

20. Hong, N., Zhang, N., Wu, H., Lu, S., Yu, Y., Hou, L., Lu, Y., Liu, H., and Jiang, G. (2018). Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. BMC Medical Informatics and Decision Making *18*, 81-88.
21. Bardenheuer, K., Van Speybroeck, M., Hague, C., Nikai, E., and Price, M. (2022). Haematology Outcomes Network in Europe (HONEUR)—A collaborative, interdisciplinary platform to harness the potential of real-world data in hematology. European Journal of Haematology *109*, 138-145.
22. Cho, J., You, S.C., Lee, S., Park, D., Park, B., Hripcsak, G., and Park, R.W. (2020). Application of epidemiological geographic information system: an open-source spatial analysis tool based on the OMOP Common Data Model. International journal of environmental research and public health *17*, 7824.
23. Unberath, P., Prokosch, H.U., Gründner, J., Erpenbeck, M., Maier, C., and Christoph, J. (2020). EHR-independent predictive decision support architecture based on OMOP. Applied clinical informatics *11*, 399-404.
24. Yu, Y., Ruddy, K.J., Wen, A., Zong, N., Tsuji, S., Chen, J., Shah, N.D., and Jiang, G. (2020). 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 on Translational Science Proceedings *2020*, 710.
25. Glicksberg, B.S., Burns, S., Currie, R., Griffin, A., Wang, Z.J., Haussler, D., Goldstein, T., and Collisson, E. (2020). Blockchain-authenticated sharing of genomic and clinical outcomes data of patients with cancer: a prospective cohort study. Journal of medical Internet research *22*, e16810.
26. Shin, S.J., You, S.C., Park, Y.R., Roh, J., Kim, J.-H., Haam, S., Reich, C.G., Blacketer, C., Son, D.-S., and Oh, S. (2019). Genomic common data model for seamless interoperation of biomedical data in clinical practice: retrospective study. Journal of medical Internet research *21*, e13249.
27. Kim, Y., Tian, Y., Yang, J., Huser, V., Jin, P., Lambert, C.G., Park, H., You, S.C., Park, R.W., and Rijnbeek, P.R. (2020). Comparative safety and effectiveness of alendronate versus raloxifene in women with osteoporosis. Scientific reports *10*, 11115.
28. Spotnitz, M.E., Natarajan, K., Ryan, P.B., and Westhoff, C.L. (2020). Relative risk of cervical neoplasms among copper and levonorgestrel-releasing intrauterine system users. Obstetrics & Gynecology *135*, 319-327.
29. You, S.C., Seo, S.I., Falconer, T., Yanover, C., Duarte-Salles, T., Seager, S., Posada, J.D., Shah, N.H., Nguyen, P.-A., and Kim, Y. (2023). Ranitidine Use and Incident Cancer in a Multinational Cohort. JAMA network open *6*, e2333495-e2333495.
30. Na, J., Zong, N., Wang, C., Midthun, D.E., Luo, Y., Yang, P., and Jiang, G. (2021). Characterizing phenotypic abnormalities associated with high-risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench. Journal of the American Medical Informatics Association *28*, 2313-2324.
31. Chen, R., Ryan, P., Natarajan, K., Falconer, T., Crew, K.D., Reich, C.G., Vashisht, R., Randhawa, G., Shah, N.H., and Hripcsak, G. (2020). Treatment patterns for chronic comorbid conditions in patients with cancer using a large-scale observational data network. JCO clinical cancer informatics *4*, 171-183.

32. Belenkaya, R., Gurley, M.J., Golozar, A., Dymshyts, D., Miller, R.T., Williams, A.E., Ratwani, S., Siapos, A., Korsik, V., and Warner, J. (2021). Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clinical Cancer Informatics *5*.
33. Kalokyri, V., Kondylakis, H., Sfakianakis, S., Nikiforaki, K., Karatzanis, I., Mazzetti, S., Tachos, N., Regge, D., Fotiadis, D.I., and Marias, K. (2023). MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes. JCO Clinical Cancer Informatics *7*, e2300101.
34. Jiang, X., Beaton, M.A., Gillberg, J., Williams, A., and Natarajan, K. (2022). Feasibility of Linking Area Deprivation Index Data to the OMOP Common Data Model. (American Medical Informatics Association), pp. 587.
35. Park, J., Lee, J.Y., Moon, M.H., Park, Y.H., and Rho, M.J. (2023). Cancer research line (CAREL): development of expanded distributed research networks for prostate cancer and lung cancer. Technology in Cancer Research & Treatment *22*, 15330338221149262.
36. Yoo, S., Yoon, E., Boo, D., Kim, B., Kim, S., Paeng, J.C., Yoo, I.R., Choi, I.Y., Kim, K., and Ryoo, H.G. (2022). Transforming thyroid cancer diagnosis and staging information from unstructured reports to the observational medical outcome partnership common data model. Applied Clinical Informatics *13*, 521-531.
37. Park, J., You, S.C., Jeong, E., Weng, C., Park, D., Roh, J., Lee, D.Y., Cheong, J.Y., Choi, J.W., and Kang, M. (2021). A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study. JMIR medical informatics *9*, e23983.
38. Ryu, B., Yoon, E., Kim, S., Lee, S., Baek, H., Yi, S., Na, H.Y., Kim, J.-W., Baek, R.-M., and Hwang, H. (2020). Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. Journal of medical Internet research *22*, e18526.
39. Gruendner, J., Schwachhofer, T., Sippl, P., Wolf, N., Erpenbeck, M., Gulden, C., Kapsner, L.A., Zierk, J., Mate, S., and Stürzl, M. (2019). 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.
40. Lee, S.-H., Chun, K.J., Park, J., Kim, J., Sung, J.D., Park, R.W., Choi, J., and Yang, K. (2021). Angiotensin converting enzyme inhibitors and incidence of lung cancer in a population based cohort of common data model in Korea. Scientific Reports *11*, 18576.
41. Lee, J.-H., Kim, S., Kim, K., Chai, Y.J., Yu, H.W., Kim, S.-J., Choi, J.Y., Chung, Y.S., Lee, K.E., and Yi, K.H. (2021). Assessment of inter-institutional post-operative hypoparathyroidism status using a common data model. Journal of Clinical Medicine *10*, 4454.
42. Seol, S., Choi, J.R., Choi, B., Kim, S., Jeon, J.Y., Park, K.N., Park, J.H., Park, M.W., Eun, Y.-G., and Park, J.J. (2023). Effect of statin use on head and neck cancer prognosis in a multicenter study using a Common Data Model. Scientific Reports *13*, 19770.
43. Lee, S.-H., Park, J., Park, R.W., Shin, S.J., Kim, J., Sung, J.D., Kim, D.J., and Yang, K. (2022). Renin-angiotensin-aldosterone system inhibitors and risk of Cancer: a Population-Based Cohort Study using a Common Data Model. Diagnostics *12*, 263.
44. Lee, Y.H., Kim, D.-H., Kim, J., and Lee, J. (2022). Risk assessment of postoperative pneumonia in cancer patients using a common data model. Cancers *14*, 5988.

45. Kim, S., Bang, J.-I., Boo, D., Kim, B., Choi, I.Y., Ko, S., Yoo, I.R., Kim, K., Kim, J., and Joo, Y. (2022). Second primary malignancy risk in thyroid cancer and matched patients with and without radioiodine therapy analysis from the observational health data sciences and informatics. European Journal of Nuclear Medicine and Molecular Imaging *49*, 3547-3556.
46. Ha, H., Ko, Y.-H., Kim, K., Hong, J., Lee, G.-W., Jeong, S.H., Bang, S.-M., and Yoon, S.-S. (2023). Application of the Khorana score for cancer-associated thrombosis prediction in patients of East Asian ethnicity undergoing ambulatory chemotherapy. Thrombosis Journal *21*, 63.
47. Felmeister, A.S., Waanders, A.J., Leary, S.E., Stevens, J., Mason, J.L., Teneralli, R., Hu, X., and Bailey, L.C. (2017). Preliminary exploratory data analysis of simulated national clinical data research network for future use in annotation of a rare tumor biobanking initiative. (IEEE), pp. 2098-2104.
48. Meystre, S.M., Heider, P.M., Kim, Y., Aruch, D.B., and Britten, C.D. (2019). Automatic trial eligibility surveillance based on unstructured clinical data. International journal of medical informatics *129*, 13-19.
49. Lee, A.R., Park, H., Yoo, A., Kim, S., Sunwoo, L., and Yoo, S. (2023). Risk prediction of Emergency Department visits in patients with Lung Cancer using machine learning: Retrospective Observational Study. JMIR Medical Informatics *11*, e53058.
50. Bräuner, K.B., Rosen, A.W., Tsouchnika, A., Walbech, J.S., Gögenur, M., Lin, V.A., Clausen, J.S., and Gögenur, I. (2022). Developing prediction models for short-term mortality after surgery for colorectal cancer using a Danish national quality assurance database. International Journal of Colorectal Disease *37*, 1835-1843.
51. Hartwig, M., Bräuner, K.B., Vogelsang, R., and Gögenur, I. (2022). Preoperative prediction of lymph node status in patients with colorectal cancer. Developing a predictive model using machine learning. International Journal of Colorectal Disease *37*, 2517-2524.
52. Lin, V., Tsouchnika, A., Allakhverdiiev, E., Rosen, A., Gögenur, M., Clausen, J., Bräuner, K., Walbech, J., Rijnbeek, P., and Drakos, I. (2022). Training prediction models for individual risk assessment of postoperative complications after surgery for colorectal cancer. Techniques in Coloproctology *26*, 665-675.
53. Seneviratne, M.G., Banda, J.M., Brooks, J.D., Shah, N.H., and Hernandez-Boussard, T.M. (2018). Identifying cases of metastatic prostate cancer using machine learning on electronic health records. (American Medical Informatics Association), pp. 1498.
54. Yoon, J.Y., Kwak, M.S., Kim, H.I., and Cha, J.M. (2021). Seasonal variations in the diagnosis of the top 10 cancers in Korea: a nationwide population-based study using a common data model. Journal of Gastroenterology and Hepatology *36*, 3371-3380.
55. Seo, S.I., Park, C.H., Kim, T.J., Bang, C.S., Kim, J.Y., Lee, K.J., Kim, J., Kim, H.H., You, S.C., and Shin, W.G. (2022). 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 Medicine *11*, 1217-1231.
56. Kim, T., Seo, S.I., Lee, K.J., Park, C.H., Kim, T.J., Kim, J., and Shin, W.G. (2022). Decreasing incidence of gastric cancer with increasing time after helicobacter pylori treatment: a nationwide population-based cohort study. Antibiotics *11*, 1052.
57. Seo, S.I., Kim, T.J., Park, C.H., Bang, C.S., Lee, K.J., Kim, J., Kim, H.H., and Shin, W.G. (2022). Incidence and survival outcomes of colorectal cancer in long-term

metformin users with diabetes: a population-based cohort study using a common data model. Journal of Personalized Medicine *12*, 584.

58. Seo, S.I., Park, C.H., You, S.C., Kim, J.Y., Lee, K.J., Kim, J., Kim, Y., Yoo, J.J., Seo, W.-W., and Lee, H.S. (2021). 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.
59. Maier, C., Lang, L., Storf, H., Vormstein, P., Bieber, R., Bernarding, J., Herrmann, T., Haverkamp, C., Horki, P., and Laufer, J. (2018). Towards implementation of OMOP in a German university hospital consortium. Applied clinical informatics *9*, 054-061.
60. Song, Q., Bates, B., Shao, Y.R., Hsu, F.-C., Liu, F., Madhira, V., Mitra, A.K., Bergquist, T., Kavuluru, R., and Li, X. (2022). Risk and outcome of breakthrough COVID-19 infections in vaccinated patients with cancer: real-world evidence from the National COVID Cohort Collaborative. Journal of Clinical Oncology *40*, 1414.
61. Carus, J., Nürnberg, S., Ückert, F., Schlüter, C., and Bartels, S. (2022). Mapping cancer registry data to the episode domain of the Observational Medical Outcomes Partnership Model (OMOP). Applied Sciences *12*, 4010.
62. Carus, J., Trübe, L., Szczepanski, P., Nürnberg, S., Hees, H., Bartels, S., Nennecke, A., Ückert, F., and Gundler, C. (2023). Mapping the Oncological Basis Dataset to the Standardized Vocabularies of a Common Data Model: A Feasibility Study. Cancers *15*, 4059.
63. Michael, C.L., Sholle, E.T., Wulff, R.T., Roboz, G.J., and Campion Jr, T.R. (2020). Mapping local biospecimen records to the OMOP common data model. AMIA Summits on Translational Science Proceedings *2020*, 422.
64. Kondylakis, H., Kalokyri, V., Sfakianakis, S., Marias, K., Tsiknakis, M., Jimenez-Pastor, A., Camacho-Ramos, E., Blanquer, I., Segrelles, J.D., and López-Huguet, S. (2023). Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. European Radiology Experimental *7*, 20.
65. Warner, J.L., Dymshyts, D., Reich, C.G., Gurley, M.J., Hochheiser, H., Moldwin, Z.H., Belenkaya, R., Williams, A.E., and Yang, P.C. (2019). HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. Journal of biomedical informatics *96*, 103239.
66. Hong, N., Zhang, N., Wu, H., Lu, S., Yu, Y., Hou, L., Lu, Y., Liu, H., and Jiang, G. (2018). Preliminary exploration of survival analysis using the OHDSI common data model: a case study of intrahepatic cholangiocarcinoma. BMC Medical Informatics and Decision Making *18*, 81-88.
67. OHDSI. OHDSI Publications. https://www.ohdsi.org/publications/
68. Wang, L., Fu, S., Wen, A., Ruan, X., He, H., Liu, S., Moon, S., Mai, M., Riaz, I.B., and Wang, N. (2022). Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. JCO Clinical Cancer Informatics *6*, e2200006.
69. Liu, S., Wen, A., Wang, L., He, H., Fu, S., Miller, R., Williams, A., Harris, D., Kavuluru, R., and Liu, M. (2023). An open natural language processing (NLP) framework for EHR-based clinical research: a case demonstration using the National COVID Cohort Collaborative (N3C). Journal of the American Medical Informatics Association *30*, 2036-2040.

70. Wen A, W.L., He H, Fu S, Liu S, Hanauer DA, Harris DR, Kavuluru R, Zhang R, Natarajan K, Pavinkurve NP, Hajagos J, Rajupet S, Lingam V, Saltz M, Elowsky C, Moffitt RA, Koraishy FM, Palchuk MB, Donovan J, Lingrey L, Stone-DerHargopian G, Miller RT, Williams AE, Leese PJ, Kovach PI, Pfaff ER, Zemmel M, Pates RD, Guthe N, Haendel MA, Chute CG, Liu H, National COVID Cohort Collaborative , the RECOVER Initiative (2024). An NLP System for COVID/PASC: A Case Demonstration of the OHNLP Toolkit from the National COVID Cohort Collaborative and the RECOVER programs. JMIR Medical Informatics *forthcoming/in press*.
71. Lee, Y.W., Strong, D.M., Kahn, B.K., and Wang, R.Y. (2002). AIMQ: a methodology for information quality assessment. Information & management *40*, 133-146.
72. Fu, S., Wen, A., Pagali, S., Zong, N., St Sauver, J., Sohn, S., Fan, J., and Liu, H. (2022). The implication of latent information quality to the reproducibility of secondary use of electronic health records. Studies in health technology and informatics *290*, 173.
73. NLM Value Set Authority Center https://vsac.nlm.nih.gov/.
74. Peterson, K.J., Jiang, G., Brue, S.M., Shen, F., and Liu, H. (2017). Mining hierarchies and similarity clusters from value set repositories. (American Medical Informatics Association), pp. 1372.
75. Peterson, K.J., Jiang, G., Brue, S.M., and Liu, H. (2016). Leveraging terminology services for extract-transform-load processes: a user-centered approach. (American Medical Informatics Association), pp. 1010.

Appendix

Summary of studies using NLP

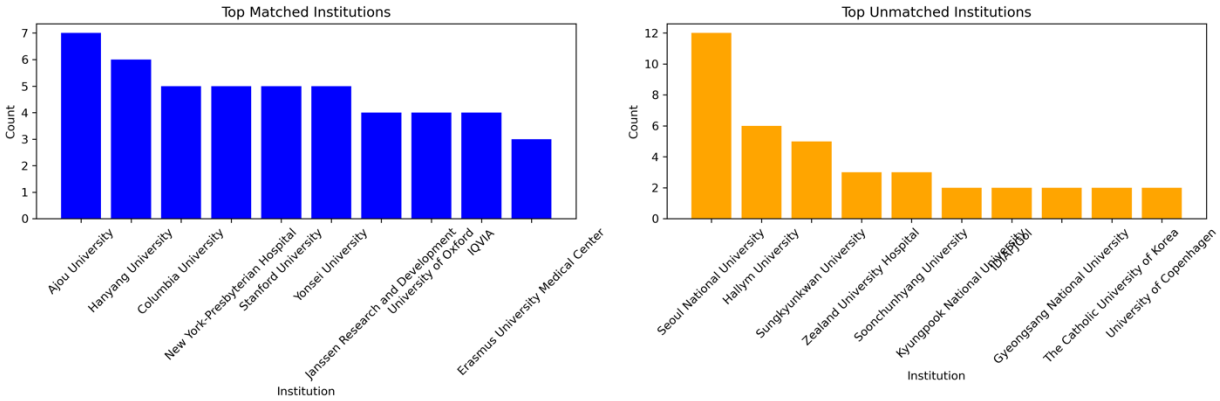
Appendix Table 1 shows the summary of papers using NLP for OMOP-based cancer studies. The data analysis theme has 1 study in the clinical trial domain. [48] As a large majority of eligibility criteria is only mentioned in unstructured clinical text besides demographics, NLP was shown to extract eligibility criteria from unstructured clinical notes with high precision and recall. OMOP was used for the computable representation of eligibility criteria. Additionally, we identified 3 papers from the "Infrastructure" theme using South Korean EHRs. [36-38] One proposed a framework for hierarchical annotation of textual data and integration into a standardized OMOP-CDM medical database. [37] This study utilized topic modeling to identify medical concepts within the unstructured documents and conducted multidimensional validation by identifying associations, such as the association of node positivity with mortality in patients with colorectal cancer. In an effort to transform pathology reports into the CDM, regular expression rules were used to extract clinical and genetic information. [38] Manual chart review was conducted for validation but no result was reported for the NLP performance. In another study, thyroid cancer diagnosis and cancer stage information were extracted from pathology reports and whole-body scan reports. [36] Of the 4 studies using NLP, 3 used rule-based methods, only 1 study worked on multi-site data from three metropolitan university hospitals, [36] and 3 conducted NLP evaluation or validation in only one institution. [48,37,38]

Appendix Table 1. A summary of papers adopting NLP

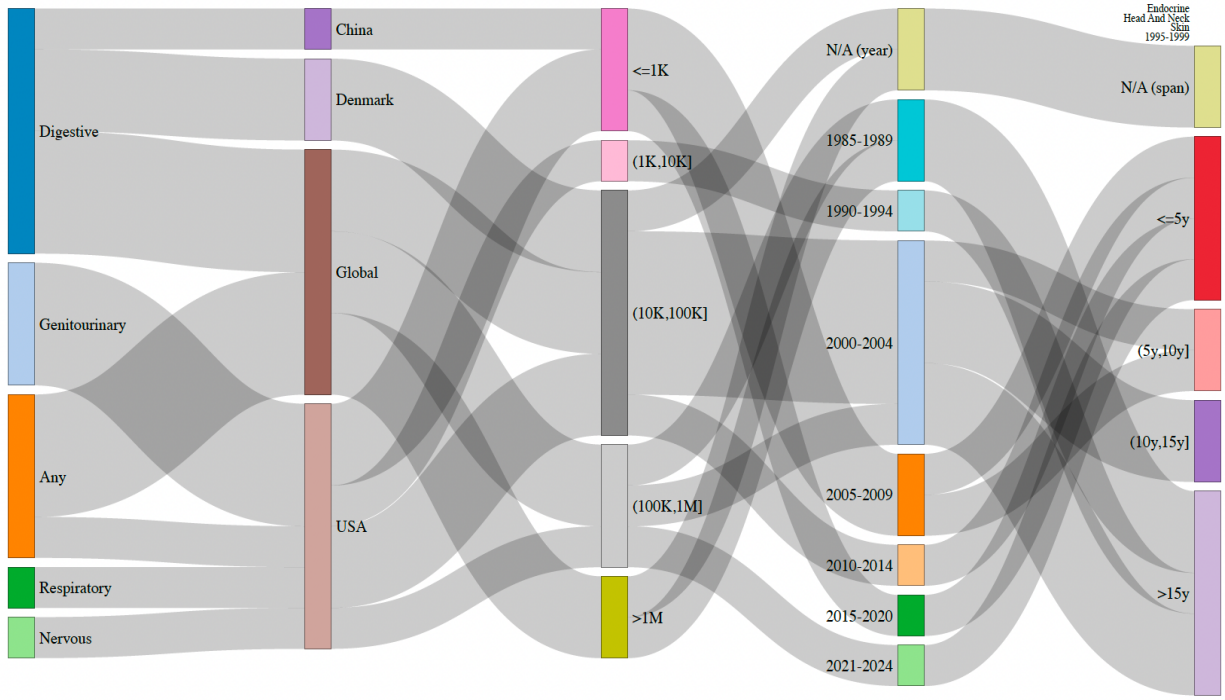
Themes	Paper	Publication year	NLP methods	Document type	Extracted data elements	Geographic area
Advanced analysis	Automatic trial eligibility surveillance ….	2019	rule-based and ML-based	clinical notes	ECOG, cancer staging, biomarkers, functional status and menopausal status	U.S.
Infrastructure	Transformation of Pathology Reports Into …	2020	rule-based	Pathology Reports	Cancer diagnosis, genetics	South Korea
	A Framework (SOCRATex) for Hierarchical Annotation …	2021	topic modeling manual annotation	Pathology reports, radiology reports, and admission notes	Not pre-defined	South Korea
	Transforming Thyroid Cancer Diagnosis and Staging Information …	2022	rule-based	pathology reports and whole-body scan reports	Cancer diagnosis and staging	South Korea

Institution names analysis

Among 26 unique studies that reported institution names, we compared the names with OHDSI collaborators (https://www.ohdsi.org/who-we-are/collaborators/). In total, we identified 92 unique institutions. Among them, 36 (38%) institutions' names were found on the collaborator list. As suggested by Figure x1, Ajou University, Hanyang University, and Columbia University are the top contributors for the in-network sites. Seoul National University, Hallym University, and Sungkyunkwan University are the top contributors for the out-network sites.



Appendix Figure 1. Distribution comparison between in-network (matched) and out-network sites.



Appendix Figure 2. Linkage between the aggregated cancer type, geographic area, cohort size, start year of study, and study period. Analysis based on countries excluding South Korea.

Appendix 1. Search Strategies

IEEE Xplore

27 articles resulted from: (cancer* OR tumor* OR tumour* OR neoplas* OR oncology*) in Abstract AND ("ohdsi"OR ("observational health data sciences Informatics") OR "omop" OR "observational medical outcomes partnership" OR "common data model") in Abstract AND (2010-2023 in Year)

Ovid

Journals@Ovid@TMC Library (subscribed full text)

Journals@Ovid (some full text)
Ovid MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations and Daily <1946 to January 12, 2024>

1	(hodgkin* NEAR 1 disease or adenocarcinoma* or adenoma* or anticarcinogen* or Astrocytoma* or blastoma* or burkitt* or cancer* or carcinogen* or carcinoid* or carcinom* or carcinosarcoma* or chordoma* or "Chronic Myeloproliferative Disorder*" or craniopharyngioma* or ependymoma* or Esthesioneuroblastoma* or germinoma* or "gestational trophoblastic disease*" or Glioblastoma* or glioma* or gonadoblastoma* or hepatoblastoma* or histeocytoma* or histiocytoma* or histiocytos* or leukaemi* or leukemi* or lymphangioma* or lymphangiomyoma* or lymphangiosarcoma* or lymphom* or Macroglobulinemia* or malignan* or melanom* or meningioma* or mesenchymoma* or mesonephroma* or Mesothelioma* or metasta* or "multiple myeloma*" or "Mycosis Fungoide*" or neoplas* or neuroblastoma* or neuroma* or nonmelanoma* or nsclc or oncogen* or oncolog* or ostesarcoma* or Papillomatos* or paraganglioma* or paraneoplas* or pheochromocytoma* or plasmacytoma* or precancerous or retinoblastoma* or Rhabdomyosarcoma* or Sarcoma* or "section 16" or "Szary Syndrome*" or teratocarcinoma* or teratoma* or tumor* or tumour*).ab. or (hodgkin* NEAR 1 disease or adenocarcinoma* or adenoma* or anticarcinogen* or Astrocytoma* or blastoma* or burkitt* or cancer* or carcinogen* or carcinoid* or carcinom* or carcinosarcoma* or chordoma* or "Chronic Myeloproliferative Disorder*" or craniopharyngioma* or ependymoma* or Esthesioneuroblastoma* or germinoma* or "gestational trophoblastic disease*" or Glioblastoma* or glioma* or gonadoblastoma* or hepatoblastoma* or histeocytoma* or histiocytoma* or histiocytos* or leukaemi* or leukemi* or lymphangioma* or lymphangiomyoma* or lymphangiosarcoma* or lymphom* or Macroglobulinemia* or malignan* or melanom* or meningioma* or mesenchymoma* or mesonephroma* or Mesothelioma* or metasta* or "multiple myeloma*" or "Mycosis Fungoide*" or neoplas* or neuroblastoma* or neuroma* or nonmelanoma* or nsclc or oncogen* or oncolog* or ostesarcoma* or Papillomatos* or paraganglioma* or paraneoplas* or pheochromocytoma* or plasmacytoma* or precancerous or retinoblastoma* or Rhabdomyosarcoma* or Sarcoma* or "section 16" or "Szary Syndrome*" or teratocarcinoma* or teratoma* or tumor* or tumour*).mp. [mp=ti, ab, tx, ct, bt, ot, nm, hw, fx, kf, ox, px, rx, ui, sy, ux, mx]	9297984
2	limit 1 to english language	8625230
3	limit 2 to full text	2071817
4	limit 3 to yr="2010 - 2023"	1140731
5	limit 4 to original articles	624061
6	(ohdsi or observational health data sciences Informatics or omop or observational medical outcomes partnership or common data model).ab. or (ohdsi or observational health data sciences Informatics or omop or observational medical outcomes	2280

	partnership or common data model).mp. [mp=ti, ab, tx, ct, bt, ot, nm, hw, fx, kf, ox, px, rx, ui, sy, ux, mx]	
7	5 and 6	112
8	remove duplicates from 7	59

PubMed

Search number	Query	Search Details	Results
3	#1 AND #2	(((("hodgkin*"[All Fields] AND "NEAR"[All Fields]) AND "1 disease"[Title/Abstract]) OR "adenocarcinoma*"[Title/Abstract] OR "adenoma*"[Title/Abstract] OR "anticarcinogen*"[Title/Abstract] OR "astrocytoma*"[Title/Abstract] OR "blastoma*"[Title/Abstract] OR "burkitt*"[Title/Abstract] OR "cancer*"[Title/Abstract] OR "carcinogen*"[Title/Abstract] OR "carcinoid*"[Title/Abstract] OR "carcinom*"[Title/Abstract] OR "carcinosarcoma*"[Title/Abstract] OR "chordoma*"[Title/Abstract] OR "chronic myeloproliferative disorder*"[Title/Abstract] OR "craniopharyngioma*"[Title/Abstract] OR "ependymoma*"[Title/Abstract] OR "esthesioneuroblastoma*"[Title/Abstract] OR "germinoma*"[Title/Abstract] OR "gestational trophoblastic disease*"[Title/Abstract] OR "glioblastoma*"[Title/Abstract] OR "glioma*"[Title/Abstract] OR "gonadoblastoma*"[Title/Abstract] OR "hepatoblastoma*"[Title/Abstract] OR "histiocytoma*"[Title/Abstract] OR "histiocytos*"[Title/Abstract] OR "leukaemi*"[Title/Abstract] OR "leukemi*"[Title/Abstract] OR "lymphangioma*"[Title/Abstract] OR "lymphangiomyoma*"[Title/Abstract] OR "lymphangiosarcoma*"[Title/Abstract] OR "lymphom*"[Title/Abstract] OR "macroglobulinemia*"[Title/Abstract] OR "malignan*"[Title/Abstract] OR "melanom*"[Title/Abstract] OR "meningioma*"[Title/Abstract] OR "mesenchymoma*"[Title/Abstract] OR	85

		"mesonephroma*"[Title/Abstract] OR "mesothelioma*"[Title/Abstract] OR "metasta*"[Title/Abstract] OR "multiple myeloma*"[Title/Abstract] OR "mycosis fungoide*"[Title/Abstract] OR "neoplas*"[Title/Abstract] OR "neuroblastoma*"[Title/Abstract] OR "neuroma*"[Title/Abstract] OR "nonmelanoma*"[Title/Abstract] OR "nsclc"[Title/Abstract] OR "oncogen*"[Title/Abstract] OR "oncolog*"[Title/Abstract] OR "ostesarcoma*"[Title/Abstract] OR "papillomatos*"[Title/Abstract] OR "paraganglioma*"[Title/Abstract] OR "paraneoplas*"[Title/Abstract] OR "pheochromocytoma*"[Title/Abstract] OR "plasmacytoma*"[Title/Abstract] OR "precancerous"[Title/Abstract] OR "retinoblastoma*"[Title/Abstract] OR "rhabdomyosarcoma*"[Title/Abstract] OR "sarcoma*"[Title/Abstract] OR "section 16"[Title/Abstract] OR "teratocarcinoma*"[Title/Abstract] OR "teratoma*"[Title/Abstract] OR "tumor*"[Title/Abstract] OR "tumour*"[Title/Abstract]) AND ((2010/01/01:2023/12/31[Date - Publication] AND ("ohdsi"[Title/Abstract] OR (("observational"[All Fields] AND ("health"[MeSH Terms] OR "health"[All Fields] OR "health s"[All Fields] OR "healthful"[All Fields] OR "healthfulness"[All Fields] OR "healths"[All Fields]) AND ("data basel"[Journal] OR "brown univ dig addict theory appl"[Journal] OR "data"[All Fields])) AND "sciences informatics"[Title/Abstract]) OR "omop"[Title/Abstract] OR "observational medical outcomes partnership"[Title/Abstract])) OR "common data model"[Title/Abstract])	
2	("2010/01/01"[Date - Publication] : "2023/12/31"[Date - Publication]) AND ((ohdsi[Title/Abstract]) OR (observational health data sciences informatics[Title/Abstract]) OR (omop[Title/Abstract]) OR (observational medical outcomes	(2010/01/01:2023/12/31[Date - Publication] AND ("ohdsi"[Title/Abstract] OR (("observational"[All Fields] AND ("health"[MeSH Terms] OR "health"[All Fields] OR "health s"[All Fields] OR "healthful"[All Fields] OR "healthfulness"[All Fields] OR "healths"[All Fields]) AND ("data	675

	partnership[Title/Abstract])) OR (common data model[Title/Abstract])	basel"[Journal] OR "brown univ dig addict theory appl"[Journal] OR "data"[All Fields])) AND "sciences informatics"[Title/Abstract]) OR "omop"[Title/Abstract] OR "observational medical outcomes partnership"[Title/Abstract])) OR "common data model"[Title/Abstract]	
3	((((hodgkin* NEAR/1 disease[Title/Abstract]) OR adenocarcinoma*[Title/Abstract] OR adenoma*[Title/Abstract] OR anticarcinogen*[Title/Abstract] OR Astrocytoma*[Title/Abstract] OR blastoma*[Title/Abstract] OR burkitt*[Title/Abstract] OR cancer*[Title/Abstract] OR carcinogen*[Title/Abstract] OR carcinoid*[Title/Abstract] OR carcinom*[Title/Abstract] OR carcinosarcoma*[Title/Abstract] OR chordoma*[Title/Abstract] OR "Chronic Myeloproliferative Disorder*"[Title/Abstract] OR craniopharyngioma*[Title/Abstract] OR ependymoma*[Title/Abstract] OR Esthesioneuroblastoma*[Title/Abstract] OR germinoma*[Title/Abstract] OR "gestational trophoblastic disease*"[Title/Abstract] OR Glioblastoma*[Title/Abstract] OR glioma*[Title/Abstract] OR gonadoblastoma*[Title/Abstract] OR hepatoblastoma*[Title/Abstract] OR histeocytoma*[Title/Abstract] OR histiocytoma*[Title/Abstract] OR histiocytos*[Title/Abstract] OR leukaemi*[Title/Abstract] OR leukemi*[Title/Abstract] OR lymphangioma*[Title/Abstract] OR lymphangiomyoma*[Title/Abstract] OR lymphangiosarcoma*[Title/Abstract] OR lymphom*[Title/Abstract] OR Macroglobulinemia*[Title/Abstract] OR malignan*[Title/Abstract] OR melanom*[Title/Abstract] OR meningioma*[Title/Abstract] OR mesenchymoma*[Title/Abstract] OR mesonephroma*[Title/Abstract] OR	(("hodgkin*"[All Fields] AND "NEAR"[All Fields]) AND "1 disease"[Title/Abstract]) OR "adenocarcinoma*"[Title/Abstract] OR "adenoma*"[Title/Abstract] OR "anticarcinogen*"[Title/Abstract] OR "astrocytoma*"[Title/Abstract] OR "blastoma*"[Title/Abstract] OR "burkitt*"[Title/Abstract] OR "cancer*"[Title/Abstract] OR "carcinogen*"[Title/Abstract] OR "carcinoid*"[Title/Abstract] OR "carcinom*"[Title/Abstract] OR "carcinosarcoma*"[Title/Abstract] OR "chordoma*"[Title/Abstract] OR "chronic myeloproliferative disorder*"[Title/Abstract] OR "craniopharyngioma*"[Title/Abstract] OR "ependymoma*"[Title/Abstract] OR "esthesioneuroblastoma*"[Title/Abstract] OR "germinoma*"[Title/Abstract] OR "gestational trophoblastic disease*"[Title/Abstract] OR "glioblastoma*"[Title/Abstract] OR "glioma*"[Title/Abstract] OR "gonadoblastoma*"[Title/Abstract] OR "hepatoblastoma*"[Title/Abstract] OR "histiocytoma*"[Title/Abstract] OR "histiocytos*"[Title/Abstract] OR "leukaemi*"[Title/Abstract] OR "leukemi*"[Title/Abstract] OR "lymphangioma*"[Title/Abstract] OR "lymphangiomyoma*"[Title/Abstract] OR "lymphangiosarcoma*"[Title/Abstract] OR "lymphom*"[Title/Abstract] OR "macroglobulinemia*"[Title/Abstract] OR "malignan*"[Title/Abstract] OR "melanom*"[Title/Abstract] OR "meningioma*"[Title/Abstract] OR "mesenchymoma*"[Title/Abstract] OR "mesonephroma*"[Title/Abstract] OR "mesothelioma*"[Title/Abstract] OR "metasta*"[Title/Abstract] OR "multiple	4,791,627

	Mesothelioma*[Title/Abstract] OR metasta*[Title/Abstract] OR "multiple myeloma*"[Title/Abstract] OR "Mycosis Fungoide*"[Title/Abstract] OR neoplas*[Title/Abstract] OR neuroblastoma*[Title/Abstract] OR neuroma*[Title/Abstract] OR nonmelanoma*[Title/Abstract] OR nsclc[Title/Abstract] OR oncogen*[Title/Abstract] OR oncolog*[Title/Abstract] OR ostesarcoma*[Title/Abstract] OR Papillomatos*[Title/Abstract] OR paraganglioma*[Title/Abstract] OR paraneoplas*[Title/Abstract] OR pheochromocytoma*[Title/Abstract] OR plasmacytoma*[Title/Abstract] OR precancerous[Title/Abstract] OR retinoblastoma*[Title/Abstract] OR Rhabdomyosarcoma*[Title/Abstract] OR Sarcoma*[Title/Abstract] OR "section 16"[Title/Abstract] OR "Szary Syndrome*"[Title/Abstract] OR teratocarcinoma*[Title/Abstract] OR teratoma*[Title/Abstract] OR tumor*[Title/Abstract] OR tumour*[Title/Abstract])))	myeloma*"[Title/Abstract] OR "mycosis fungoide*"[Title/Abstract] OR "neoplas*"[Title/Abstract] OR "neuroblastoma*"[Title/Abstract] OR "neuroma*"[Title/Abstract] OR "nonmelanoma*"[Title/Abstract] OR "nsclc"[Title/Abstract] OR "oncogen*"[Title/Abstract] OR "oncolog*"[Title/Abstract] OR "ostesarcoma*"[Title/Abstract] OR "papillomatos*"[Title/Abstract] OR "paraganglioma*"[Title/Abstract] OR "paraneoplas*"[Title/Abstract] OR "pheochromocytoma*"[Title/Abstract] OR "plasmacytoma*"[Title/Abstract] OR "precancerous"[Title/Abstract] OR "retinoblastoma*"[Title/Abstract] OR "rhabdomyosarcoma*"[Title/Abstract] OR "sarcoma*"[Title/Abstract] OR "section 16"[Title/Abstract] OR "teratocarcinoma*"[Title/Abstract] OR "teratoma*"[Title/Abstract] OR "tumor*"[Title/Abstract] OR "tumour*"[Title/Abstract]	

Web of Science

Entitlements	#	Search Query	Database	Results	Date Run
- WOS.SCI: 1900 to 2024 - WOS.AHCI: 1975 to 2024 - WOS.ESCI: 2005 to 2024 - WOS.ISTP: 1990 to 2024 - WOS.SSCI: 1900 to 2024 - WOS.ISSHP: 1990 to 2024	1	((((((((((TI=(OHDSI)) OR TI=("observational health data sciences and informatics")) OR TI=(omop)) OR TI=("observational medical outcomes partnership")) OR TI=("common data model")) OR AB=(ohdsi)) OR AB=("observational health data sciences and informatics")) OR AB=(mop)) OR AB=("observational medical outcomes partnership")) OR AB=("common data model"))) AND DOP=(2010-01-01/2023-12-31)	Web of Science Core Collection	8825	Sat Jan 13 2024 18:51:34 GMT-0600 (Central Standard Time)
- WOS.SCI: 1900 to 2024 - WOS.AHCI:	2	(TI=(hodgkin* NEAR/1 disease) OR TI=(adenocarcinoma*) OR TI=(adenoma*) OR TI=(anticarcinogen*) OR TI=(Astrocytoma*) OR	Web of Science Core Collection	5526873	Sat Jan 13 2024 18:56:33 GMT-

1975 to 2024 - WOS.ESCI: 2005 to 2024 - WOS.ISTP: 1990 to 2024 - WOS.SSCI: 1900 to 2024 - WOS.ISSHP: 1990 to 2024		TI=(blastoma*) OR TI=(burkitt*) OR TI=(cancer*) OR TI=(carcinogen*) OR TI=(carcinoid*) OR TI=(carcinom*) OR TI=(carcinosarcoma*) OR TI=(chordoma*) OR TI=("Chronic Myeloproliferative Disorder*") OR TI=(craniopharyngioma*) OR TI=(ependymoma*) OR TI=(Esthesioneuroblastoma*) OR TI=(germinoma*) OR TI=("gestational trophoblastic disease*") OR TI=(Glioblastoma*) OR TI=(glioma*) OR TI=(gonadoblastoma*) OR TI=(hepatoblastoma*) OR TI=(histeocytoma*) OR TI=(histiocytoma*) OR TI=(histiocytos*) OR TI=(leukaemi*) OR TI=(leukemi*) OR TI=(lymphangioma*) OR TI=(lymphangiomyoma*) OR TI=(lymphangiosarcoma*) OR TI=(lymphom*) OR TI=(Macroglobulinemia*) OR TI=(malignan*) OR TI=(melanom*) OR TI=(meningioma*) OR TI=(mesenchymoma*) OR TI=(mesonephroma*) OR TI=(Mesothelioma*) OR TI=(metasta*) OR TI=("multiple myeloma*") OR TI=("Mycosis Fungoide*") OR TI=(neoplas*) OR TI=(neuroblastoma*) OR TI=(neuroma*) OR TI=(nonmelanoma*) OR TI=(nsclc) OR TI=(oncogen*) OR TI=(oncolog*) OR TI=(ostesarcoma*) OR TI=(Papillomatos*) OR TI=(paraganglioma*) OR TI=(paraneoplas*) OR TI=(pheochromocytoma*) OR TI=(plasmacytoma*) OR TI=(precancerous) OR TI=(retinoblastoma*) OR TI=(Rhabdomyosarcoma*) OR TI=(Sarcoma*) OR TI=("section 16") OR TI=("Szary Syndrome*") OR TI=(teratocarcinoma*) OR TI=(teratoma*) OR TI=(tumor*) OR TI=(tumour*)) OR (AB=(hodgkin* NEAR/1 disease) OR AB=(adenocarcinoma*) OR AB=(adenoma*) OR AB=(anticarcinogen*) OR			0600 (Central Standard Time)

		AB=(Astrocytoma*) OR AB=(blastoma*) OR AB=(burkitt*) OR AB=(cancer*) OR AB=(carcinogen*) OR AB=(carcinoid*) OR AB=(carcinom*) OR AB=(carcinosarcoma*) OR AB=(chordoma*) OR AB=("Chronic Myeloproliferative Disorder*") OR AB=(craniopharyngioma*) OR AB=(ependymoma*) OR AB=(Esthesioneuroblastoma*) OR AB=(germinoma*) OR AB=("gestational trophoblastic disease*") OR AB=(Glioblastoma*) OR AB=(glioma*) OR AB=(gonadoblastoma*) OR AB=(hepatoblastoma*) OR AB=(histeocytoma*) OR AB=(histiocytoma*) OR AB=(histiocytos*) OR AB=(leukaemi*) OR AB=(leukemi*) OR AB=(lymphangioma*) OR AB=(lymphangiomyoma*) OR AB=(lymphangiosarcoma*) OR AB=(lymphom*) OR AB=(Macroglobulinemia*) OR AB=(malignan*) OR AB=(melanom*) OR AB=(meningioma*) OR AB=(mesenchymoma*) OR AB=(mesonephroma*) OR AB=(Mesothelioma*) OR AB=(metasta*) OR AB=("multiple myeloma*") OR AB=("Mycosis Fungoide*") OR AB=(neoplas*) OR AB=(neuroblastoma*) OR AB=(neuroma*) OR AB=(nonmelanoma*) OR AB=(nsclc) OR AB=(oncogen*) OR AB=(oncolog*) OR AB=(ostesarcoma*) OR AB=(Papillomatos*) OR AB=(paraganglioma*) OR AB=(paraneoplas*) OR AB=(pheochromocytoma*) OR AB=(plasmacytoma*) OR AB=(precancerous) OR AB=(retinoblastoma*) OR AB=(Rhabdomyosarcoma*) OR AB=(Sarcoma*) OR AB=("section 16") OR AB=("Szary Syndrome*") OR AB=(teratocarcinoma*) OR AB=(teratoma*) OR AB=(tumor*) OR AB=(tumour*))			
- WOS.SCI: 1900	3	#1 AND #2	Web of Science	272	Sat Jan 13 2024

to 2024 - WOS.AHCI: 1975 to 2024 - WOS.ESCI: 2005 to 2024 - WOS.ISTP: 1990 to 2024 - WOS.SSCI: 1900 to 2024 - WOS.ISSHP: 1990 to 2024			Core Collection		18:57:59 GMT-0600 (Central Standard Time)