Najia Ahmadi, Yuan Peng, Markus Wolfien et al.

OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction : A Systematic Review

Erstveröffentlichung in / First published in:

International journal of molecular sciences. 2022. 23. MDPI AG. ISSN: 1422-0067.

DOI: https://doi.org/10.3390/ijms231911834

International Journal of
Molecular Sciences

MDPI

Review

OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review

Najia Ahmadi *, Yuan Peng, Markus Wolfien, Michéle Zoch and Martin Sedlmayr

Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Fetscherstraße 74, 01307 Dresden, Germany
* Correspondence: najia.ahmadi@tu-dresden.de

Abstract: The current generation of sequencing technologies has led to significant advances in identifying novel disease-associated mutations and generated large amounts of data in a high-throughput manner. Such data in conjunction with clinical routine data are proven to be highly useful in deriving population-level and patient-level predictions, especially in the field of cancer precision medicine. However, data harmonization across multiple national and international clinical sites is an essential step for the assessment of events and outcomes associated with patients, which is currently not adequately addressed. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an internationally established research data repository introduced by the Observational Health Data Science and Informatics (OHDSI) community to overcome this issue. To address the needs of cancer research, the genomic vocabulary extension was introduced in 2020 to support the standardization of subsequent data analysis. In this review, we evaluate the current potential of the OMOP CDM to be applicable in cancer prediction and how comprehensively the genomic vocabulary extension of the OMOP can serve current needs of AI-based predictions. For this, we systematically screened the literature for articles that use the OMOP CDM in predictive analyses in cancer and investigated the underlying predictive models/tools. Interestingly, we found 248 articles, of which most use the OMOP for harmonizing their data, but only 5 make use of predictive algorithms on OMOP-based data and fulfill our criteria. The studies present multicentric investigations, in which the OMOP played an essential role in discovering and optimizing machine learning (ML)-based models. Ultimately, the use of the OMOP CDM leads to standardized data-driven studies for multiple clinical sites and enables a more solid basis utilizing, e.g., ML models that can be reused and combined in early prediction, diagnosis, and improvement of personalized cancer care and biomarker discovery.

Keywords: OHDSI; OMOP CDM; EHR; PLP; prediction; machine learning

Citation: 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.* **2022**, *23*, 11834. https://doi.org/10.3390/ijms231911834

Academic Editors: Ekaterini Chatzaki and Nima Aghaeepour

Received: 8 July 2022
Accepted: 1 October 2022
Published: 5 October 2022

1. Introduction

Electronic health record (EHR) data have been used to store patient-specific information for decades, including structured data, such as diagnosis, medication, laboratory test results, and unstructured data obtained from clinical reports. Observational patient data are used in vast computational analyses, including the generation of individual patient profiles and detection of patient similarity based on clinical and genomics data [1]. With advancements in the field of genetics, it is possible to analyze large amounts of genomic data using different ML and other predictive methods that can widen the knowledge about diseases with a genetic background, such as cancer, as well as rare and unclear diseases. However, such algorithms need large patient cohorts to reach a clinical prediction scale and useful diagnostic decision support [2]. For this purpose, a harmonized research data repository is necessary to enable a joint analysis across institutions based on observational data [3].

The Observational Health Data Science and Informatics (OHDSI) initiative is a promising international effort to optimize secondary use of observational data by harmonizing and standardizing clinical data and to create scalable analytical tools [4,5]. The basis for this is the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which ensures homogeneous storage of observational healthcare data across different databases with interoperable formats and standard terminologies. The terminologies for diagnoses/conditions, observations, and drugs within the OMOP CDM are based on, for example, International Classification of Diseases (ICD) codes [6], Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) [7], and normalized naming system for generic and branded drugs (RxNorm) [8]. To apply these concepts, one usually needs to retrieve the already mapped tables from the Automated Terminology Harmonization, Extraction and Normalization for Analysis (ATHENA) [9] standardized vocabulary tool from OHDSI. Afterwards, the harmonized data stored in the OMOP CDM format can be used in systematic studies, population-level estimations, drug and biomarker evaluations, as well as further patient-level prediction [10].

To develop tailor-made therapies for cancer patients, researchers must have access to genetic variants and their associated pathways together with the clinical information. Studies such as Unberath et al. [11] created a vocabulary set using the HUGO Gene Nomenclature Committee (HGNC) [12] for their specific use case, but a comprehensive standard vocabulary that can assist in sequencing data on the OMOP CDM and predictive modules in general would be essential.

The first attempt for this purpose was a Genomic Common Data Model (G-CDM) proposed in 2019 [13] to store next-generation sequencing (NGS) data. The G-CDM introduced four extension tables to the OMOP CDM which, to use any of the standardized OHDSI tools for the purpose of data analysis or prediction on data stored using this format, would require adoption of the tools, because they strictly follow the OMOP CDM structure. One important tool is named ATLAS – A unified interface for the OHDSI tools [14], which is an open-source web-based interface to configure analysis modules such as Patient Level Prediction (PLP) in the OMOP CDM [15,16]. In particular, PLP can be used to define Artificial Intelligence (AI)-based analyses on previously defined patient cohorts by using an easy-to-use graphical user interface. However, to tackle the challenge of enabling oncology data in the OMOP CDM without any structural changes, the OHDSI community has published the first version of a set of new vocabularies for presenting genomic data in the OMOP CDM in 2020 [17]. These new vocabularies are built based on different gene databases, e.g., ClinVar [18], Clinical Interpretation of Variants in Cancer (CIVic) [19], and Precision Oncology Knowledge Base (OncoKB) [20]. With the use of new vocabularies, the data can be represented in the OMOP CDM without the need for any further extensions. Additionally, an OMOP oncology module was introduced in 2021 [3], which extends the OMOP CDM and its terminologies to support the harmonized documentation of cancer conditions, treatment, and disease abstraction. This module uses the concepts from seven existing standards, namely, International Classification of Diseases for Oncology (ICD-O-3) [21], Hematology Oncology (HemOnc) [22,23], North American Association of Central Cancer Registries (NAACCR) [24], College of American Pathologists CAP [25], Nebraska Lexicon [26], National Cancer Institute (NCI) [27], and Anatomical Therapeutic Chemical (ATC) classification [28]. Moreover, the Radiology Common Data Model (R-CDM) for standardization of Digital Imaging Communications in Medicine (DICOM) was published in 2021 [29]. R-CDM uses the RadLex glossary, which contains 75,000 radiology terms to harmonize DICOM imaging data into two extended tables, radiology occurrence and radiology image, on the OMOP CDM.

Clinical integration of the OMOP CDM can pave the way back to patients through facilitating access to relevant data, enabling multicentric, multidatabase studies to enhance statistical power and transfer results across populations [30–32]. With the recent advancements in the field of medical informatics, many predictive algorithms are known and used in the field of oncology, which benefits largely from the use of such models in uncovering unknown information about the cause and course of certain types of cancer. For the purpose of this review, we looked for studies that have taken advantage of such predictive models to perform cancer-related analyses on an OMOP CDM and evaluate to what extent the genomic vocabulary extension of the OMOP can serve current needs of ML-based predictions.

Research Questions

Since this review aims to evaluate the applicability and current potential of the OMOP CDM in cancer prediction, the research questions that guided this study are as follows:

1. To what level are predictive models used and integrated in cancer prediction on the OMOP CDM, and where does it require more attention?
2. Does the existing genomic vocabulary in the OMOP cover the needs for analysis of genomic data using, e.g., the ATLAS PLP module?
3. What tools (other than PLP) exist to support these predictive analyses on cancer data in the OMOP?

2. Results

The literature screening resulted in 248 papers from 13 search engines, of which only five matched the scope of our review and are finally included (Figure 1). In particular, after duplicates' removal, 212 articles were title- and abstract-screened. In this step, articles that did not indicate an AI-based prediction analysis and OMOP in their title or abstract were excluded.

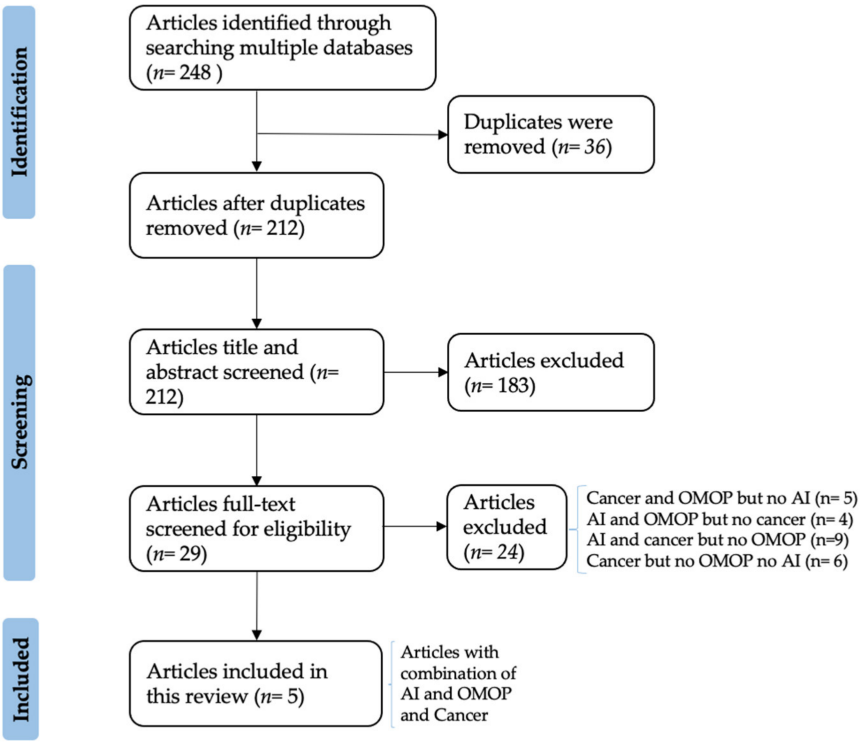
Identification
Articles identified through searching multiple databases (*n*= 248)
Duplicates were removed (*n*= 36)
Articles after duplicates removed (*n*= 212)
Screening
Articles title and abstract screened (*n*= 212)
Articles excluded (*n*= 183)
Articles full-text screened for eligibility (*n*= 29)
Articles excluded (*n*= 24)
Cancer and OMOP but no AI (n= 5)
AI and OMOP but no cancer (n= 4)
AI and cancer but no OMOP (n=9)
Cancer but no OMOP no AI (n= 6)
Included
Articles included in this review (*n*= 5)
Articles with combination of AI and OMOP and Cancer

Figure 1. PRISMA Flow-chart diagram showing the paper selection process.

Afterwards, the full text screening step analyzed in total 29 articles, out of which 15 were either focused on AI and cancer but without using the OMOP or focused solely on cancer or AI. The remaining nine articles either contained cancer studies on OMOP-based data not using predictive AI models [11,33–36] or performed predictive analysis on OMOP-based data of a non-cancerous disease [37–45]. An example for the first group are preliminary studies that are focused on harmonizing data in the OMOP using extract, load, and transform (ETL) processes. The articles that perform predictive analysis on other than cancerous data partially use different machine learning and deep learning methods. One of these studies is Hardin et al. [46] that uses the OHDSI PLP module for the development of predictive models. Since these excluded studies also contain a valuable source of information for the current review, detailed information of the most important excluded articles and the finally included five articles can be obtained in the attached Supplementary Table S1 (color-coded in grey). In the following, we highlight the studies that ultimately contain aspects of AI and the OMOP in the cancer domain.

Among the included papers, Felmeister et al. [1] focus on the pediatric rare brain tumor and follow an exploratory approach to extract pertinent information from a large simulated observational dataset based on the OMOP and discover data points that contribute to the data-driven phenotype of a diagnosed subject. An example of such a data point is population-based survival estimates. The authors apply a supervised prediction approach and take advantage of the Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Neighbors classifier (KNN), Decision Tree classifier (CART), Gaussian Naïve Bayes (NB), and Support Vector Machine (SVM) algorithms. The models are applied on a simulated cohort of 1000, in which KNN performs best with the highest percentage of correctly identified cases. SVM and LR are the second and third best-performing ML algorithms. The analysis shows that usage of OMOP CDM observational data in exploration analysis can lead to valuable discoveries.

Meystre et al. [47] train a Natural Language Processing (NLP) method using manually annotated physician letters for subsequent automatic detection of patient eligibility for breast cancer clinical trials. The authors encode the clinical trial eligibility criteria to the corresponding clinical information system. The clinical notes were stored in the notes table in the OMOP CDM. They use NLP and an SVM classifier method to extract the patient-derived EHR information from the existing free text notes written by physicians. The cohort was designed by using the ATLAS platform of OHDSI. The study shows that NLP is able to extract the eligibility criteria for clinical trials from EHR notes from a cohort of 229 patients, with an average recall and precision of 84.6% and 64.4%. In comparison, SVM models perform better with an average recall of 90.9% and precision of 89.7%. Using the extracted eligibility information, the patients were classified to determine eligibility using an SVM binary classifier with high accuracy.

Unlike Felmeister et al., the third study by Seneviratne et al. [48] uses tree-based classification models, such as Lasso Penalty (LASSO), Random Forest (RF), Gradient Boosted Machine (GBM), and Extreme Gradient Boosting (XGB), on a cohort with prostate cancer. The algorithms classify metastatic cancer from non-metastatic cases based on the stage of cancer, which is usually documented in text form in medical notes, which means it is only feasible to extract cancer stage information on population level, when an AI-based approach is used. The study demonstrates identification of patients with metastatic prostate cancer in a cohort of 5861 patients using an RF classifier with a precision and recall of 90% and 40%, respectively. The RF model outperforms other models, including normal ICD code search, which leads to a recall and precision of 54% and 33%.

Moreover, the Information Technology for the Future of Cancer (ITFoC) [49] introduces a framework for the validation of AI algorithms with omics and clinical data for prediction of the treatment response in triple-negative breast cancer (TNBC) [50]. In this framework, the AI models will be developed and validated on real-world data. The clinical and -omics data will be harmonized via the OMOP CDM and terminologies, such as ICD-10, Logical Observation Identifiers Names and Codes (LOINC), and SNOMED-CT.

Furthermore, Lee et al. [51] perform a retrospective study of data obtained from seven hospitals in Korea that adopted the OMOP CDM as main research data repository. The study aims to find the association of angiotensin-converting enzyme inhibitor (ACEi) and angiotensin receptor blocker (ARB) with lung cancer development. Similar to Meystre et al., for cohort definition and defining the baseline characteristics of the study, the OHDSI tool ATLAS was used.

As shown in Table 1, all of the abovementioned articles use the OMOP CDM as a data standardization model, transform their datasets to this format, and design their AI-based analysis on it. The vocabularies that are used in these papers to transform data into the OMOP CDM structure include the International Classification of Diseases Clinical Modification, 9th Revision (ICD-9-CM), International Classification of Diseases Clinical Modification, 10th Revision (ICD-10-CM), SNOMED-CT, and LOINC codes. Moreover, a wide range of models are used as predictive models in the aforementioned papers, starting from classical machine learning methods, e.g., RF, GBM, all the way to other regression and classification methods, including linear regression, lasso regression, SVM, and k-Nearest Neighbors (KNN). Since the OMOP CDM harmonizes different data structures, the same predictive tool or trained model can be applied in different medical studies.

Table 1. An overview of the dataset size and features used in the articles, vocabularies used to transform them into OMOP CDM format, and predictive models used to analyze the data.

Article	Dataset Size	Features	Vocabularies	Predictive Models
Felmeister et al. 2017 [1]	1000 Patients	patient, condition, observation, drug exposure and demographics (gender, race, date of birth, etc.)	ICD-9-CM, ICD-10-CM, SNOMED-CT	LR, LDA, KNN, CART, NB, and SVM
Meystre et al. 2019 [47]	229 Patients	patient identifier, gender, date of birth, height, weight, diagnostic code, procedure code, and clinical notes	LOINC, SNOMED-CT	NLP and SVM
Seneviratne et al. 2018 [48]	5861 Patients	conditions, procedures, medications, observations, and laboratory values	ICD-9 and ICD-10	LASSO, RF, GBM, and XGB
Tsopra et al. 2021 [50]	-	-	ICD-10, LOINC, and SNOMED-CT	-
Lee et al. 2021 [51]	207,794 Patients	age group, medical history: general (e.g., dementia, cardiovascular disease (e.g., arterial fibrillation), and neoplasms (e.g., malignant neoplasm of anorectum)	-	Cox regression

In summary, all of the abovementioned finally included articles use the OMOP as a common data model, out of which two (Meystre et al. and Lee et al.) use ATLAS for the purpose of cohort definition. Seneviratne et al. use only tree-based methods, whereas the other studies commonly utilize a combination of different types of methods (e.g., tree-based, boosting, SVMs). The use of different methods can also be obtained by using the PLP from OHDSI. Only Felmeister et al. use Centers for Medicare and Medicaid (CMS) Medicare Claims Synthetic Public Use Files (SynPUF) simulated data [52], which is a freely available dataset converted to the OMOP CDM used for benchmarking studies and technology implementations. A single study (Meystre et al.) uses unstructured, i.e., free text data for the initial analysis, and Tsopra et al. uses -omics data in addition to structured clinical data.

3. Discussion

In the era of cancer precision medicine, the need for targeted treatment protocols is increasing, and only predictive analyses may provide comprehensive information [53–55].

The genomic vocabulary in the OMOP CDM is a step forward towards harmonization of genomic data, which has the potential to enable analyses on combined clinical and sequencing data. However, three of the articles included in this review were published before 2020, meaning that they do not use the genomic vocabularies in the OMOP CDM but rather limit their features to clinical data. Lee et al. 2021 [51] do not mention a use of genomic vocabulary, but in the framework introduced by Tsopra et al. [50], the clinical and -omics data for TNBC will be harmonized on the OMOP CDM. Therefore, although the OMOP has enabled successful large-scale AI-based studies for some of the included articles, so far, there is no explicit evidence of how predictive methods can be applied to real-world oncological data represented in the OMOP CDM via genomics vocabularies. Moreover, a use of PLP is recommended by the OHDSI community also for genomic data analyses because this way, it conforms with standard practices. Currently, PLP is not mentioned in any of the articles that are in this review and focuses on cancer analysis on the OMOP. PLP is designed to support the clinical decision-making process based on the available individual medical history of the patient. The methods suggested in the articles in Table 1 are tested on the dataset by the authors and are not evaluated in routine care. Interestingly, most of the AI algorithms used in the articles (e.g., RF, KNN, and SVM) are included in the current version of PLP (https://github.com/OHDSI/PatientLevelPrediction/tree/main/R (accessed on 7 July 2022). Additionally, the PLP module also supports a set of deep learning methods, such as convolutional neural networks and recurrent neural networks, but we did not observe these methods in the included studies. This is an indication that most of the existing analyses might be replaced by the standard OHDSI tool to answer the underlying research questions [15,16]. This would improve the reusability and transparency of multicentric studies in general. However, it is also possible to customize AI-based prediction analyses for OMOP CDMs with individual programming languages and packages in R or Python, which allows for a more individualized analysis, if desired. An indication of the low numbers of published manuscripts that are currently available could be attributed to the ongoing multiple standardization efforts within OHDSI, which are highly dependent on currently developed infrastructures and implementations, such as PLP and genomic vocabulary of the OMOP.

It is worth mentioning that while our search string suffices as regards our research question (OMOP CDM's ability concerning cancer prediction), it did not include the words "AI" and "oncology", and a broader search string, not limited to only "cancer", may possibly lead to different results. Interestingly, there are multiple international consortia arising built upon the OMOP. In particular, the Harmony Alliance recently presented a prime example of a European-wide OMOP-CDM-based study, in which an ML-based tool was introduced to predict the risk of relapse after first remission in leukemia patients [31]. The Data Analytics and Real World Interrogation Network (DARWIN EU) is aiming to deliver timely and reliable evidence by use of OMOP-based real-world data on disease and patient population, and the use, safety, and effectiveness of medicines, including vaccines, throughout the lifecycle of a medicinal product [56,57]. Moreover, the European Network of Excellence for Big Data in Prostate Cancer (PIONEER) performs predictive analysis on OMOP-based patient data [58,59]. These large-scale research studies likewise demonstrate the current need and versatile application scenarios for more envisaged OMOP-CDM-based studies in the future.

4. Materials and Methods

This paper follows the guidelines of the PRISMA extension for scoping reviews (PRISMA-ScR) [60]. According to the research questions stated above, we have the following search string: "cancer AND ((machine learning) OR (prediction) OR (algorithm)) AND (OHDSI) OR (OMOP)".

4.1. Paper Identification

We conducted a systematic literature search for articles published between 2016 and 2021 in the most relevant databases in the domain, namely, PubMed, BMC, JAMIA, Journal of Bioinformatics, PLOS ONE, Hindawi, BMC Medical Informatics and Decision Making, Elsevier, Sage, Springer, Science Direct, Nature, and IEEE, with our search string. The article search was conducted on 27 April 2022. The results of our article search are processed within a library in the Zotero Citation Manager [61].

4.2. Paper Inclusion and Exclusion Criteria

The inclusion and exclusion criteria used in this study are shown in Table 2.

Table 2. Inclusion and exclusion criteria for the title and abstract screening and full-text screening.

Screening Round	Inclusion	Exclusion
Title and abstract screening	The article is primary research in a peer-reviewed journal or conference.	The article is of any other type, for instance, study protocols, commentaries and editorials, tutorials, project reports, medical case studies, and master and doctoral thesis.
	The article is written in English.	The article is written in a language other than English.
	The title or abstract mention analysis of cancer data.	The title or abstract do not mention analysis of cancer data.
	The title or abstract mention OMOP or OHDSI.	The title or abstract do not mention OMOP or OHDSI.
Full-text screening	The article allows open access to full text.	The article does not allow open access to full text.
	The article defines a predictive approach for cancer medicine.	The article defines a predictive approach but for a domain other than cancer medicine
	The predictive approach in the article uses the OMOP CDM as the data model.	The predictive approach in the article does not use the OMOP CDM as the data model.

4.3. Selection and Review of Articles

After duplicate removal by the built-in function of Zotero, the study selection process was performed in two steps in accordance with the criteria from the Table 2**.** Two authors (NA, YP) independently performed a title and abstract screening and discussed the conflicts. In the second step, both authors (NA, YP) performed a full-text screening on the articles and resolved the conflicts after discussion.

4.4. Data Charting

After identifying relevant articles, an extraction table with a focus on the use of the OMOP CDM, the utilized predictive methods, and the terminologies used was created by NA and approved by YP. The data extraction was carried out by NA and approved by YP.

5. Conclusions

In line with the recently published perspective by Rehm et al. 2021 [62], we see the implementation of standards and frameworks for clinical data sharing as a critical step to advancing genomic medicine. In particular, the OMOP CDM has the potential to enable international collaborative analyses, which is a significant component for cancer precision medicine. Using the comprehensive genomics vocabularies, oncology data can be harmonized in the OMOP CDM, and this can lead to advancements in this field. Moreover, given that these vocabularies enable mapping and transferring oncology data into the OMOP CDM, the analysis of those data on a population level is now more possible than ever before.

Since the genomic vocabulary on the OMOP CDM and the R-CDM was recently published, these can be seen as a first, initial step that paves the way for mapping and transferring oncology data into the OMOP CDM. However, most of the articles included in this review were published before the release of the genomic vocabularies for the OMOP CDM. Therefore, a follow-up study needs to be conducted to obtain a measurable impact about how comprehensively the genomic vocabularies add up to existing AI models' accuracy within the field of cancer precision medicine.

In addition, a broad use and continuous development of standards, especially within the field of oncology research, is essential to fully utilize the benefits of data harmonization across different cancer care entities. The use of AI-based approaches, such as NLP, SVM, and RF, has proven efficient in identification/classification of patients with certain characteristics and led to an improvement in the cohort identification process for clinical trials and other observational research [47,48]. In brief, standardized data, represented, for example, as the OMOP CDM, can serve as a solid base to enable a decentralized use of AI models that is needed for an optimized analysis on population-level estimation, patient-level prediction, and more specifically in cancer survival, time-based analysis, and biomarker discovery [1].

This review provides a first hallmark into current applications and the usefulness of the OMOP CDM for AI-based cancer prediction and likewise summarizes and promotes the beneficial use of OMOP CDMs in cancer prediction. In prospective future works, we will evaluate the application of PLP in ATLAS, as a predictive AI framework, for the purpose of cancer precision medicine in the OMOP CDM with the genomic vocabularies [17]. This focus, which is beyond the current study, will investigate whether existing modules in PLP are able to handle real-world oncological datasets using genomics vocabularies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms231911834/s1.

Author Contributions: N.A.: conceptualization, N.A., Y.P.: methodology, data acquisition, literature screening, N.A.: data analysis and interpretation, NA: writing—original draft preparation, N.A., M.W., Y.P., M.Z., M.S.: writing—review and editing, N.A.: visualization, M.S.: resources, all authors take responsibility for the scientific integrity of the work. All authors have read and agreed to the published version of the manuscript.This study was performed by NA to (partially) fulfill the requirements for obtaining the aca-demic degree "Dr. rer. medic." from the Technische Universität Dresden.

Funding: This work is accomplished as part of the SATURN project funded by the Federal Ministry of Health (BMG), FKZ: ZMI1-2520DAT02C, and the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative; MIRACUM Consortium, FKZ: 01lZZ1801L (Dresden).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This study was performed by NA to (partially) fulfill the requirements for obtaining the academic degree ″Dr. rer. medic.“ from the Technische Universität Dresden.

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

References

1. Felmeister, A.S.; Waanders, A.J.; Leary, S.E.S.; Stevens, J.; Mason, J.L.; Teneralli, R.; Hu, X.; Bailey, L.C. Preliminary exploratory data analysis of simulated national clinical data research network for future use in annotation of a rare tumor biobanking initiative. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; IEEE: Kansas City, MO, USA, 2017; pp. 2098–2104.
2. Briganti, G.; Le Moine, O. Artificial Intelligence in Medicine: Today and Tomorrow. *Front. Med.* **2020**, *7*, 27. [CrossRef] [PubMed]
3. Belenkaya, R.; Gurley, M.J.; Golozar, A.; Dymshyts, D.; Miller, R.T.; Williams, A.E.; Ratwani, S.; Siapos, A.; Korsik, V.; Warner, J.; et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin. Cancer Inform.* **2021**, *5*, 12–20. [CrossRef] [PubMed]
4. 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.; Rijnbeek, P.R.; et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud. Health Technol. Inform.* **2015**, *216*, 574–578.

5. Garza, M.; Del Fiol, G.; Tenenbaum, J.; Walden, A.; Zozus, M.N. Evaluating common data models for use with a longitudinal community registry. *J. Biomed. Inform.* **2016**, *64*, 333–341. [CrossRef] [PubMed]
6. ICD. Available online: https://www.dimdi.de/dynamic/de/klassifikationen/icd/ (accessed on 30 June 2022).
7. Millar, J. The Need for a Global Language-SNOMED CT Introduction. *Stud. Health Technol. Inform.* **2016**, *225*, 683–685.
8. RxNorm. Available online: https://www.nlm.nih.gov/research/umls/rxnorm/index.html (accessed on 30 June 2022).
9. Athena. Available online: https://athena.ohdsi.org/search-terms/start (accessed on 30 June 2022).
10. Voss, E.A.; Makadia, R.; Matcho, A.; Ma, Q.; Knoll, C.; Schuemie, M.; DeFalco, F.J.; Londhe, A.; Zhu, V.; Ryan, P.B. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 553–564. [CrossRef] [PubMed]
11. Unberath, P.; Prokosch, H.U.; Gründner, J.; Erpenbeck, M.; Maier, C.; Christoph, J. EHR-Independent Predictive Decision Support Architecture Based on OMOP. *Appl. Clin. Inform.* **2020**, *11*, 399–404. [CrossRef]
12. Povey, S.; Lovering, R.; Bruford, E.; Wright, M.; Lush, M.; Wain, H. The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.* **2001**, *109*, 678–680. [CrossRef]
13. Shin, S.J.; You, S.C.; Park, Y.R.; Roh, J.; Kim, J.-H.; Haam, S.; Reich, C.G.; Blacketer, C.; Son, D.-S.; Oh, S.; et al. Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J. Med. Internet Res.* **2019**, *21*, e13249. [CrossRef]
14. ATLAS–A unified interface for the OHDSI tools–OHDSI. Available online: https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools/ (accessed on 7 July 2022).
15. Reps, J.M.; Schuemie, M.J.; Suchard, M.A.; Ryan, P.B.; Rijnbeek, P.R. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 969–975. [CrossRef]
16. Rijnbeek, P.; Reps, J. Chapter 13 Patient-Level Prediction. In *The Book of OHDSI*; 2021. Available online: https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html (accessed on 7 July 2022).
17. Kaduk, D.; Komar, V.; Golozar, A.; Robinson, P.; Wagner, A.H.; Gurley, M.; You, S.C.; Seojeong, S.; Mehr, S.; Williams, A.; et al. Genomic Data Harmonization through the OMOP Standardized Vocabularies. In Proceedings of the 2020 OHDSI Global Symposium (Showcase 13), Virtual Symposium, 18–21 October 2020.
18. ClinVar. Available online: https://www.ncbi.nlm.nih.gov/clinvar/ (accessed on 30 June 2022).
19. CIViC-Clinical Interpretation of Variants in Cancer. Available online: https://civicdb.org/welcome (accessed on 30 June 2022).
20. OncoKBTM-MSK's Precision Oncology Knowledge Base. Available online: https://www.oncokb.org/ (accessed on 30 June 2022).
21. Percy, C.; van Holten, V.; Muir, C.S.; World Health Organization. *International Classification of Diseases for Oncology*; World Health Organization: Lyon, France, 1976.
22. Warner, J.L.; Dymshyts, D.; Reich, C.G.; Gurley, M.J.; Hochheiser, H.; Moldwin, Z.H.; Belenkaya, R.; Williams, A.E.; Yang, P.C. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *J. Biomed. Inform.* **2019**, *96*, 103239. [CrossRef] [PubMed]
23. HemOnc.org-A Hematology Oncology Wiki. Available online: https://hemonc.org/wiki/Main_Page (accessed on 17 May 2022).
24. NAACCR. Available online: https://www.naaccr.org/ (accessed on 17 May 2022).
25. Renshaw, A.A.; Mena-Allauca, M.; Gould, E.W.; Sirintrapun, S.J. Synoptic Reporting: Evidence-Based Review and Future Directions. *JCO Clin. Cancer Inform.* **2018**, *2*, 1–9. [CrossRef] [PubMed]
26. Campbell, W.S.; Campbell, J.R.; West, W.W.; McClay, J.C.; Hinrichs, S.H. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 885–892. [CrossRef] [PubMed]
27. NCI Thesaurus. Available online: https://ncithesaurus.nci.nih.gov/ncitbrowser/ (accessed on 17 May 2022).
28. Anatomical Therapeutic Chemical (ATC) Classification. Available online: https://www.who.int/tools/atc-ddd-toolkit/atc-classification (accessed on 17 May 2022).
29. Park, C.; You, S.C.; Jeon, H.; Jeong, C.W.; Choi, J.W.; Park, R.W. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Med. J.* **2022**, *63*, S74–S83. [CrossRef]
30. Bathelt, F.; Reinecke, I.; Peng, Y.; Henke, E.; Weidner, J.; Bartos, M.; Gött, R.; Waltemath, D.; Engelmann, K.; Schwarz, P.E.; et al. Opportunities of Digital Infrastructures for Disease Management—Exemplified on COVID-19-Related Change in Diagnosis Counts for Diabetes-Related Eye Diseases. *Nutrients* **2022**, *14*, 2016. [CrossRef]
31. Sobas, M.; Martinez Elicegui, J.; Villaverde Ramiro, A.; González, T.; Hernandez-Sanchez, A.; Azibeiro Melchor, R.; Benner, A.; Sträng, E.; Gastone, C.; Heckman, C.A.; et al. Harmony Alliance Provides a Machine Learning Researching Tool to Predict the Risk of Relapse after First Remission in AML Patients Treated without Allogeneic Haematopoietic Stem Cell Transplantation. *Blood* **2021**, *138*, 4041. [CrossRef]
32. Yoo, S.; Yoon, E.; Boo, D.; Kim, B.; Kim, S.; Paeng, J.C.; Yoo, I.R.; Choi, I.Y.; Kim, K.; Ryoo, H.G.; et al. Transforming Thyroid Cancer Diagnosis and Staging Information from Unstructured Reports to the Observational Medical Outcome Partnership Common Data Model. *Appl. Clin. Inform.* **2022**, *13*, 521–531. [CrossRef]
33. Liu, H.; Chi, Y.; Butler, A.; Sun, Y.; Weng, C. A knowledge base of clinical trial eligibility criteria. *J. Biomed. Inform.* **2021**, *117*, 103771. [CrossRef]

34. Jeon, H.; You, S.C.; Kang, S.Y.; Seo, S.I.; Warner, J.L.; Belenkaya, R.; Park, R.W. Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer Using Harmonized Observational Databases: Retrospective Study. *JMIR Med. Inform.* **2021**, *9*, e25035. [CrossRef]
35. Schuemie, M.J.; Weinstein, R.; Ryan, P.B.; Berlin, J.A. Quantifying bias in epidemiologic studies evaluating the association between acetaminophen use and cancer. *Regul. Toxicol. Pharmacol.* **2021**, *120*, 104866. [CrossRef]
36. Wood, W.A.; Marks, P.; Plovnick, R.M.; Hewitt, K.; Neuberg, D.S.; Walters, S.; Dolan, B.K.; Tucker, E.A.; Abrams, C.S.; Thompson, A.A.; et al. ASH Research Collaborative: A real-world data infrastructure to support real-world evidence development and learning healthcare systems in hematology. *Blood Adv.* **2021**, *5*, 5429–5438. [CrossRef]
37. Felmeister, A.S.; Waanders, A.J.; Mason, J.L.; Stevens, J.; Bailey, L.C.; Ganesan, S.; Helbig, I. A Framework for Analysis, Ontological Evaluation, and Visualization in Preparation to Predictive Analytics in Pediatric Brain Tumor Research. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea, 16–19 December 2020; pp. 1943–1950.
38. Li, J.; Tian, Y.; Zhu, Y.; Zhou, T.; Li, J.; Ding, K.; Li, J. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. *Artif. Intell. Med.* **2020**, *103*, 101814. [CrossRef]
39. Clarke, C.L.; Feigelson, H.S. Developing an Algorithm to Identify History of Cancer Using Electronic Medical Records. *eGEMs (Gener. Evid. Methods Improv. Patient Outcomes)* **2016**, *4*, 5. [CrossRef]
40. Li, J.; Tian, Y.; Li, R.; Zhou, T.; Li, J.; Ding, K.; Li, J. Improving prediction for medical institution with limited patient data: Leveraging hospital-specific data based on multicenter collaborative research network. *Artif. Intell. Med.* **2021**, *113*, 102024. [CrossRef]
41. Kim, C.; You, S.C.; Reps, J.M.; Cheong, J.Y.; Park, R.W. Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1098–1107. [CrossRef]
42. Tian, Y.; Shang, Y.; Tong, D.-Y.; Chi, S.-Q.; Li, J.; Kong, X.-X.; Ding, K.-F.; Li, J.-S. POPCORN: A web service for individual PrognOsis prediction based on multi-center clinical data CollabORatioN without patient-level data sharing. *J. Biomed. Inform.* **2018**, *86*, 1–14. [CrossRef]
43. Choi, S.; Kim, I.; Kim, Y.; Seok, J. Vocabulary Domain Prediction for Pathological Report Analysis Using ICD-O3. In Proceedings of the 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Zagreb, Croatia, 2–5 July 2019; pp. 520–522.
44. Boehm, K.M.; Khosravi, P.; Vanguri, R.; Gao, J.; Shah, S.P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **2022**, *22*, 114–126. [CrossRef] [PubMed]
45. Edmondson, M.J.; Luo, C.; Duan, R.; Maltenfort, M.; Chen, Z.; Locke, K.; Shults, J.; Bian, J.; Ryan, P.B.; Forrest, C.B.; et al. An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes. *Sci. Rep.* **2021**, *11*, 19647. [CrossRef] [PubMed]
46. Hardin, J.; Reps, J.M. Evaluating the impact of covariate lookback times on performance of patient-level prediction models. *BMC Med. Res. Methodol.* **2021**, *21*, 180. [CrossRef] [PubMed]
47. 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.* **2019**, *129*, 13–19. [CrossRef] [PubMed]
48. Seneviratne, M.G.; Banda, J.M.; Brooks, J.D.; Shah, N.H.; Hernandez-Boussard, T.M. Identifying Cases of Metastatic Prostate Cancer Using Machine Learning on Electronic Health Records. *AMIA Annu. Symp. Proc.* **2018**, *2018*, 1498–1504. [PubMed]
49. IT Future of Cancer. Available online: https://itfoc.eu/ (accessed on 17 May 2022).
50. Tsopra, R.; Fernandez, X.; Luchinat, C.; Alberghina, L.; Lehrach, H.; Vanoni, M.; Dreher, F.; Sezerman, O.U.; Cuggia, M.; de Tayrac, M.; et al. A framework for validating AI in precision medicine: Considerations from the European ITFoC consortium. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 274. [CrossRef] [PubMed]
51. Lee, S.-H.; Chun, K.J.; Park, J.; Kim, J.; Sung, J.D.; Park, R.W.; Choi, J.; Yang, K. Angiotensin converting enzyme inhibitors and incidence of lung cancer in a population based cohort of common data model in Korea. *Sci. Rep.* **2021**, *11*, 18576. [CrossRef] [PubMed]
52. CMS 2008–2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF) | CMS. Available online: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF (accessed on 19 September 2022).
53. Cirillo, D.; Valencia, A. Big data analytics for personalized medicine. *Curr. Opin. Biotechnol.* **2019**, *58*, 161–167. [CrossRef]
54. Weissler, E.H.; Naumann, T.; Andersson, T.; Ranganath, R.; Elemento, O.; Luo, Y.; Freitag, D.F.; Benoit, J.; Hughes, M.C.; Khan, F.; et al. The role of machine learning in clinical research: Transforming the future of evidence generation. *Trials* **2021**, *22*, 537. [CrossRef]
55. Chowdhury, R.I.; Tomal, J.H. Risk prediction for repeated measures health outcomes: A divide and recombine framework. *Inform. Med. Unlocked* **2022**, *28*, 100847. [CrossRef]
56. EMA Data Analysis and Real World Interrogation Network (DARWIN EU). Available online: https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu (accessed on 2 June 2022).
57. DARWIN EU Initiative Presentation–OHDSI. Available online: https://ohdsi.org/darwin-eu-initiative-presentation/ (accessed on 30 June 2022).
58. Omar, M.I.; Roobol, M.J.; Ribal, M.J.; Abbott, T.; Agapow, P.-M.; Araujo, S.; Asiimwe, A.; Auffray, C.; Balaur, I.; Beyer, K.; et al. Introducing PIONEER: A project to harness big data in prostate cancer research. *Nat. Rev. Urol.* **2020**, *17*, 351–362. [CrossRef]

59. PIONEER joins forces with EHDEN & OHDSI for prostate cancer study-a-thon. *PIONEER-Eur. Netw. Excell. Big Data Prostate Cancer 2021*. Available online: https://prostate-pioneer.eu/pioneer-joins-forces-with-ehden-ohdsi-for-prostate-cancer-study-a-thon/ (accessed on 30 June 2022).
60. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef]
61. Zotero | Your Personal Research Assistant. Available online: https://www.zotero.org/ (accessed on 10 February 2022).
62. Rehm, H.L.; Page, A.J.H.; Smith, L.; Adams, J.B.; Alterovitz, G.; Babb, L.J.; Barkley, M.P.; Baudis, M.; Beauvais, M.J.S.; Beck, T.; et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom.* **2021**, *1*, 100029. [CrossRef]