applied sciences

MDPI

*Article*

# Mapping Cancer Registry Data to the Episode Domain of the Observational Medical Outcomes Partnership Model (OMOP)

**Jasmin Carus [1,2,*], Sylvia Nürnberg [2], Frank Ückert [2], Catarina Schlüter [1] and Stefan Bartels [1,*]**

[1] University Cancer Center Hamburg (UCCH), University Hospital Hamburg-Eppendorf (UKE), 20251 Hamburg, Germany; ca.schlueter@uke.de
[2] Institute for Applied Medical Informatics, University Hospital Hamburg-Eppendorf (UKE), 20251 Hamburg, Germany; s.nuernberg@uke.de (S.N.); f.ueckert@uke.de (F.Ü.)
* Correspondence: j.carus@uke.de (J.C.); st.bartels@uke.de (S.B.)

**Abstract:** A great challenge in the use of standardized cancer registry data is deriving reliable, evidence-based results from large amounts of data. A solution could be its mapping to a common data model such as OMOP, which represents knowledge in a unified semantic base, enabling decentralized analysis. The recently released Episode Domain of the OMOP CDM allows episodic modelling of a patient' disease and treatment phases. In this study, we mapped oncology registry data to the Episode Domain. A total of 184,718 Episodes could be implemented, with the Concept of Cancer Drug Treatment most frequently. Additionally, source data were mapped to new terminologies as part of the release. It was possible to map ≈ 73.8% of the source data to the respective OMOP standard. Best mapping was achieved in the Procedure Domain with 98.7%. To evaluate the implementation, the survival probabilities of the CDM and source system were calculated ($n$ = 2756/2902, median OAS = 82.2/91.1 months, 95% CI = 77.4–89.5/84.4–100.9). In conclusion, the new release of the CDM increased its applicability, especially in observational cancer research. Regarding the mapping, a higher score could be achieved if terminologies which are frequently used in Europe are included in the Standardized Vocabulary Metadata Repository.

**Keywords:** cancer registry; standardized vocabulary; semantic interoperability; translational cancer research; common data model; OMOP; fair DATA

check for updates

**Citation:** Carus, J.; Nürnberg, S.; Ückert, F.; Schlüter, C.; Bartels, S. Mapping Cancer Registry Data to the Episode Domain of the Observational Medical Outcomes Partnership Model (OMOP). *Appl. Sci.* **2022**, *12*, 4010. https://doi.org/10.3390/app12084010

Academic Editors: Toralf Kirsten and Oya Beyan

Received: 25 February 2022
Accepted: 13 April 2022
Published: 15 April 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

A cancer diagnosis is often followed by complex treatment that can last for years. Recently, many new therapeutic approaches have been developed, either derived from basic research or the use of new diagnostic measures, such as DNA sequencing, which examine the tumor in more detail [1]. Thus, an initial diagnosis of cancer is often accompanied by a series of diagnostic modifiers, such as Gleason score, grading, stage group. From this set of characteristics, the treatment strategy can be derived, and success can be estimated. When a new therapeutic approach is selected, the physician considers which therapeutic measures have already been carried out to increase the probability of a positive response and to reduce the risk of an adverse reaction [2]. These developments in predictive medicine have ensured that guideline-based treatment is increasingly shifting towards a personalized approach. However, the ability to give a more detailed specification of the tumor phenotype due to greater stratification possibilities also leads to decreasing case numbers within a specific tumor entity. The use of an appropriate study population to achieve significant results is limited by complex inclusion and exclusion criteria. In addition, the methods for analyzing these complex relationships, for example using Artificial Intelligence (AI) techniques, are continuing to evolve. These models require larger sample sizes than the current statistical methods in order to derive valid results. The potential applications of AI in the field of oncology have grown rapidly in recent years. Especially, the use of AI in the field of image analysis has delivered great progress [3]. The identification

of complex patterns in radiological images aids the detection of malignant tumors and simplifies clinical decision-making processes. For example, one study has shown that an algorithm can predict whether a pulmonary nodule will become cancerous within the next 2 years, with 80% accuracy [4]. In addition to this use of AI in early cancer detection, image analysis can assist in the identification of tumor-specific diagnostic factors. In another study, a deep learning algorithm was used to predict the Gleason score of a patient tumor using prostatectomy images with 70% accuracy [5]. Besides the use of image analysis, AI can also help with the analysis of genomic sequencing data. As sequencing capabilities are increasing, so is the number of discovered genomic mutations, leading to researchers having to clarify associations between genomic mutations and phenotypes using literature research. This is where AI approaches might be able to simplify human workloads [6].

In modern cancer research, it is crucial to establish data exchange or decentralized analysis pipelines based on a homogeneous data semantic base in the joint networks of individual research institutions [7]. For example, Cancer Core Europe, a consortium of 28 European cancer institutions, has stated that there is a "need for creating a uniform platform for translational cancer research to bring together enough centers to generate the critical mass of patients, expertise and resource required to make a significant breakthrough in cancer care" [7] (p. 523). However, the German Cancer Consortium has identified several challenges for the establishment of such networks. Because of different data protection laws worldwide, merging data is challenging. Furthermore, depending on the data infrastructure, there are different technical requirements, such as documentation systems and others, that can make data exchange difficult. However, in general, the greatest challenge lies in semantic heterogeneity [8]. Semantic heterogeneity in this context means that two IT systems fulfill the prerequisites for receiving data from each other (syntactic interoperability), but the interpretation of this data is not possible due to ambiguous semantics. A solution could be the mapping of cancer data to a common data model (CDM) which represents knowledge with unified semantics and enables decentralized analysis. Many CDMs come with analytical applications. Thus, the integration of heterogenous operational databases into a CDM enables the use of CDM-developed analytical applications, such as package libraries and REST APIs. A well-known CDM in the field of clinical research is the PCORnet Model from the Patient-Centered Outcomes Research Institute (PCORI). They have developed a policy of data standards to enable the efficient use of data in clinical and patient-powered research without violating data protection regulations. These data standards lead to the semantic alignment of the source data, so that multi-centered studies are possible without the respective institutions having to give up control of their data [9]. This can enable larger cohort sizes, which can be analyzed using AI through federated learning. The Clinical Data Interchange Standards Consortium (CDISC) has developed several data models that cover the different phases of the clinical research process. There are data models for study planning, data collection, the tabulation of study data, and analysis. These data models maintain compatible standards across all converted datasets. In a related study using resident registry data, the most common CDMs in the clinical research domain (SCDM v.5.0, PCORnet v.3.0, OMOP v.3.0, CDISC, SDTM v.1.4.) were evaluated in terms of completeness, integrity, flexibility, integrability, and implementability for EHR-based longitudinal registry data [10]. It was found that the OMOP CDM v.3.0, provided by the Observational Health Data Sciences and Informatics (OHDSI) community, achieved the best scores regarding the evaluation criteria of the study. OHDSI is a multi-stakeholder interdisciplinary collaboration founded in 2014. It arose from the public–private partnership with the US Food and Drug Administration (FDA). After FDA funding ceased, it was decided that a collaboration should be developed; the CDM was adopted as an open-source project with the aim of integration into scientific applications. Nowadays, this collaboration consists of an international network of researchers and over 100 observational health databases from 19 countries. It develops technical solutions for the representation of uniform medical data from different source systems, tackling the lack of standardized her and EMR data and the absence of consistent patient-level data in obser-

vational databases [11]. It provides open-source applications with the goal of strengthening the research community, whose findings can then be considered in clinical questions. For example, there is a comprehensive R package library that allows feature extraction from OMOP, and AI-based analysis of these extracted OMOP data to be performed. Also of note is the PatientLevelPrediction package, which provides patient-specific prediction models using machine learning and deep learning algorithms [12]. In addition, federated pipelines of different semantic homogeneity databases reduce capture bias, and a large number of observed patients in a study leads to higher statistical power and greater stratification possibilities. In September 2021, OMOP was supplemented by a new Episode Domain [13]. This Episode Domain contains the master table Episode, which displays an episodic modeling of the course of a disease, depending on its respective concepts. Episodic modelling of cancer is essential to represent the complex disease process. Correct episodic modeling is therefore of particular importance to derive evidence from oncology data. By implementing the standardized concept of Disease Dynamic in the Episode Domain, survival probabilities with cancer-specific endpoints can be calculated via the CDM. This concept is based on the Response Evaluation Criteria in Solid Tumors (RECIST). These models were defined by an international working group aiming to establish uniform regulations for physicians to classify responses to tumor treatment [14]. The availability of RECIST data is essential to enable the comparison of the analysis results across institutions, for example from survival analyses in multicentered studies. The Episode Domain also contains the Episode_Event table which allows linkage of the abstracted Episodes to low-level events of the CDM, newly embedded with the implemented standardized vocabulary. Besides extending the CDM with the Episode Domain, new oncology terminologies, primarily those commonly used in cancer care such ICD-O-3, ATC, and others, were added to OHDSI's Standardized Vocabulary Metadata Repository.

However, the extent to which Episodes of a cancer course can be represented through the implementation of newly added tables, and how well oncology registry data can be displayed through the newly standardized vocabulary, such as ATC, HemOnc, ICDO3, and Cancer Modifier, have never been investigated. The data used in this study were collected from the clinical cancer registry (KKR) of the University Hospital Hamburg-Eppendorf (UKE), and range from the structured recording of a new diagnosis until the death of the patient within the UKE. The KKR has existed since 2010, and documents all cancer patients who have received cancer-related diagnostic or therapeutic measures at the UKE. Moreover, the KKR must report these collected cancer data to the national registry for quality assurance and research purposes.

The objective of this study was firstly to find out to what extent the source terminologies of the clinical cancer registry can be mapped to the respective OMOP standard. Secondly, we investigated to what extent the source data of the tumor documentation system can be transferred to the Episode Domain. Finally, we explored how well survival analyses can be derived from the OMOP CDM compared to the source system. Thus, overall survival analyses were conducted across the CDM and source system.

## 2. Materials and Methods

The implementation of the new tables was carried out in three phases. The first phase comprised episodic modeling according to Disease Extent, Disease Dynamic and Treatment (Figure 1). The second phase involved the mapping of the cancer data to the oncology standardized vocabulary, and the last phase comprised the linkage of Episodes to the underlying clinical events of the CDM by the implementation of the Episode_Event table.

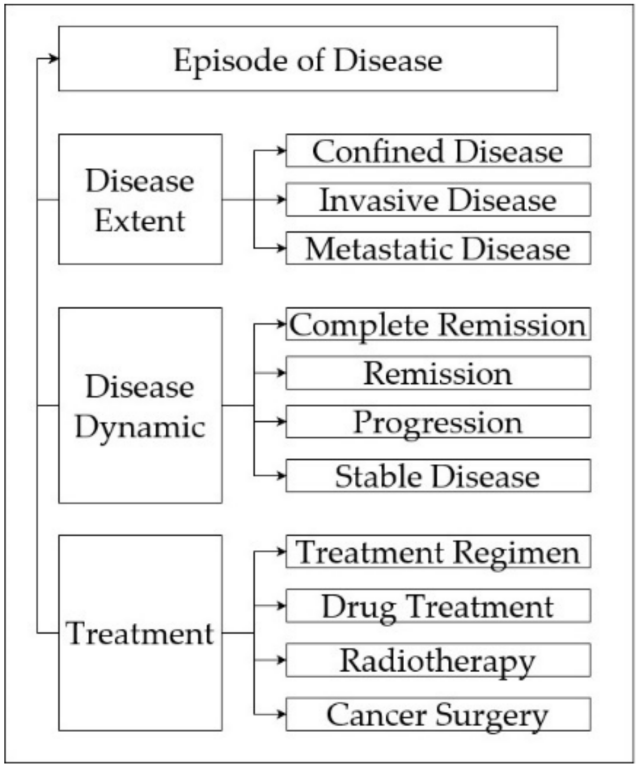



**Figure 1.** Imputed concepts involved for the episodic modeling.

### *2.1. Source System*

The Giessener Tumor Documentation System (GTDS) [15] is a client–server application with an ORACLE database management system at the backend. The frontend is provided by an ORACLE forms- and a web application. Its interface is connected to the central Health Information System (HIS) so that patients with a cancer diagnosis are automatically imported into the documentation system. All imported cases are reviewed by a clinical documentation specialist and then documented in a structured form using different input masks. The relational GTDS database comprises 422 tables that are related by primary and foreign keys. For a correct data query, a deep understanding of the cardinality of the tables is essential. Primary and foreign keys must be connected correctly to avoid either an endless query loop or duplicate data entries. A patient population of 26,000 was included in this study. This provided the groundwork for the mapping process.

### *2.2. Episodic Modelling*

In the first step, only Episodes which described the extent of the disease were modelled. Possible attributes were Confined Disease (Concept_id: 32942), Invasive Disease (Concept_id: 32943) and Metastatic Disease (Concept_id: 32944). For the modeling of these Episodes, values from the Tumor–Node–Metastases staging system (TNM) from the source system were chosen as the starting point for the modeling. Date-exact TNMs were aggregated using a custom algorithm, and time intervals were derived from these aggregated data. In addition, the source data were mapped to standardized concepts of disease response during treatment (Disease Dynamic) according to RECIST, which reflects the phase of the patient's disease and derives survival probabilities. The source system provides the disease state of the patient at a certain time point (to a day). The measurement points for the determination of the remission status are summarized in time intervals (start date, end date) under the application of a custom algorithm that firstly derives a time interval from the measurement points, secondly takes into account the underlying concept (Complete Remission (Concept_id: 32946), Remission (Concept_id: 32945), Partial Remission (Concept_id: 32947), Stable Disease (Concept_id: 32948) and Progression (Concept_id: 32949)),

and thirdly merges time intervals with the same underlying concept. For the presentation of Cancer Drug Treatment (Concept_id: 32941), Cancer Radiotherapy (Concept_id: 32940) and Cancer Surgery (Concept_id: 32939), the corresponding tables from the source system were used. The OHDSI OncoRegimenFinder algorithm [16]) was used for modeling the Treatment Regimen (Concept_id: 32531). Drugs that were administered within a 30-day time window were summarized in regimens and translated to HemOnc [17] terminology where possible.

### 2.3. Vocabulary Mapping

ICD-O3 is a combined classification of the topography and morphology of a tumor [18]. The topography is derived, in part, from the ICD-10 code and has a 4-digit character that covers the range from C00.0 to C80.9 which, similar to ICD-10, specifies the tumor category. The Morphology code of ICD-O3 specifies the type of cell of the neoplasm and the behavior. The ICD-O3 is implemented in the CDM in the Condition Domain and links the Condition_occurrence events with the disease episodes of the oncology module.

The North Association of Central Cancer Registries (NAACCR) defines cancer registry standards for the structured acquisition of data in North America [19]. NAACCR incorporates existing ontologies and classifications, such as ICDO-3, into its data standards. This ontology is mainly used in cancer registries in the USA and Canada. All data collected in the context of cancer therapy and diagnosis are assigned to specific items, which are either superordinate or assigned to special schemes, according to the respective cancer entity. Each item has a NAACCR value. Source items were mapped according to NAACR at item and value level.

The National Library of Medicine (NLM), which is part of the National Institutes of Health (NIH) of the USA, provides information and research services for making biomedical data usable in the context of healthcare, and grants access to evidence-based results [20]. In 2003, the NLM developed and administered the ontology SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms) [21]. It has nine hierarchically arranged concepts, of which this study uses Clinical Finding and Procedure, covering hierarchical levels 1 and 2. By incorporating the root concept, the underlying subtypes can be identified with their associated descendants. The higher the concept class of the corresponding domain, the more descendants can be identified in SNOMED-CT. However, it is also possible to infer the root concept from the descendants.

HemOnc is a medical Wiktionary. It provides information on treatment regimens, subdivided by disease subtypes, and additionally offers information on drugs, interventions, and general information on the treatment of neoplasms [22]. The HemOnc Wiki was integrated into the Standardized Vocabulary Metadata Repository v5 to provide a link between the abstraction of Treatment Regimen Episodes of the oncology module and low-level drug events of the OMOP CDM [23].

As part of the Episode implementation, source data were mapped to the new vocabulary (Figure 2, Table 1). In the OMOP CDM, the ICDO-3 classification was used to represent the cancer diagnosis. The elements that were used to specify the tumor diseases in more detail were included in the domains of Measurement and Observation. As part of the implementation of the Episode Domain, the source data was mapped to ICDO-3, SNOMED, ATC, HemOnc, Cancer Modifier, and NAACCR standardized vocabularies. Thereby, the primary approach was to map the oncological data to the SNOMED-CT terminology. If another classification system was more granular, with respect to cancer representation, it was preferred to SNOMED-CT.

**Table 1.** Mapped Items by Domain and Vocabulary.

| Domain | Vocabulary | Version | Items |
|---|---|---|---|
| Underlying Observation Events | SNOMED-CT | 31 July 2020 SNOMED CT International Edition; 1 August 2020 SNOMED CT US Edition; 28 October 2020 SNOMED CT UK Edition | ECOG, Histology |
| | NAACCR | NAACCR v18 | Primary Site, histology, behavior |
| Underlying Measurement Events | Cancer Modifier | Cancer Modifier 20201014 | topography, metastasis–topography, grading, lymph nodes, other classification (Gleason score, Fuhrman, WHO-ISUP, Durie and Salmon, Clark level, Masaoka staging) |
| | NAACCR | NAACCR v18 | Tumor board, regional nodes, metastasis, pathological grade, c/p TNM, c/p stage group, residual classification, Her2 |
| | SNOMED-CT | 31 July 2020 SNOMED CT International Edition; 1 August 2020 SNOMED CT US Edition; 28 October 2020 SNOMED CT UK Edition | morphology, Ann Arbor Classification, estregone/progesterone Receptor, tumor size (mm) |
| Underlying Diagnosis Events | ICDO-3 | ICDO3 SEER Site/Histology Released 06/2020 | Diagnosis |
| Abstracted Episodes | HemOnc | 26 January 2021 HemOnc | Treatment Regimen |
| | OPS | OPS Version 2020 | Cancer Surgeries |
| | ATC | 4 May 2020 ATC | Drugs |
| | SNOMED-CT | 31 July 2020 SNOMED CT International Edition; 1 August 2020 SNOMED CT US Edition; 28 October 2020 SNOMED CT UK Edition | Radiotherapy |

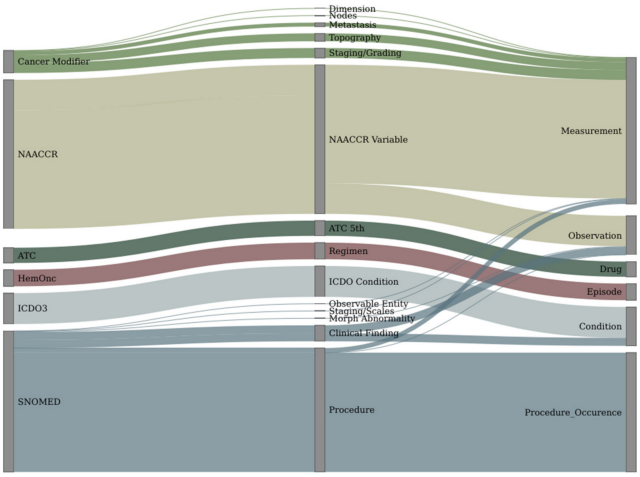



**Figure 2.** Sankey Diagram showing a flowchart of mapped vocabulary dependent on their concept class and Domain.

### *2.4. Linkage between Episodes and Underlying Clinical Events*

The linkage of the abstracted Episodes to the corresponding underlying events of the CDM was performed via the Episode_Event table. The pooling of polymorphic foreign keys of the clinical tables in the Episode_Event table provides the possibility to link the unique identifiers of low-level events of the CDM with an Episode. Thus, all therapeutic or diagnostic measures can be assigned to an Episode. For example, this table can then be used to query which measures have been undertaken during the event of a progressive course, e.g., for renewed diagnostics, re-radiation, surgery, and other measures.

### *2.5. CDM Application and Comparison*

To test the applicability and accuracy of the CDM, overall survival of a breast cancer cohort was calculated via the CDM and source system and compared to the real results. The Null Hypothesis (H0) was tested, which assumed that the calculated overall survival of the two systems was the same. The Alternative Hypothesis (H1), on the other hand, assumed that there were differences in overall survival between the systems. The probability of error (alpha error) was set at 5% for this test. This means that, if $p > 0.05$, H1 would be rejected and the H0 hypothesis could be accepted. The calculations of the survival analyses were performed in a dynamic R Markdown report. The DatabaseConnector package was used to extract the survival cohorts from the Source and CDM database. These cohorts were merged into one dataset using the dplyr package and then stratified analyzed using the Survminer and Survival packages.

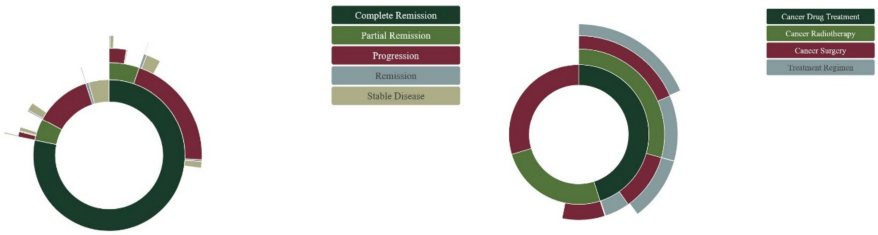
## 3. Results

### *3.1. Episodic Modeling*

Within this study, a total of 184,718 Episodes could be implemented in the new Episode table of the CDM. This standardized data pool of concepts can be used by most of the OHDSI collaborative applications, allowing cross-institutional comparison. In the Episode Domain, the concept classes of Disease Extent, Disease Dynamic and Treatment were implemented. There were 26,700 documented TNMs. From these, 18,561 could be mapped to the Disease Extent concept during modeling. Regarding the Disease Dynamic, which reflects the disease status, a total of 31,627 Disease Dynamic concepts could be derived from a total of 147,816 measurement points. The concept of Complete Remission, with 60% ($n$ = 18,980), was the most frequent outcome (Figure 3a).

With respect to the episodic modeling of treatment phases, 99,840 Treatment Episodes could be derived from the source system (Figure 3b). Within the concept class of Treatment, the Cancer Surgery could be mapped in 100% of cases (CDM: $n$ = 28,718, GTDS: $n$ = 28,718), Cancer Radiotherapy in 92.36% of cases (CDM: $n$ = 16,116, GTDS: $n$ = 17,450), Cancer Drug

Treatment in 86.32% of cases (CDM: $n$ = 37,537, GTDS: $n$ = 43,488), and Treatment Regimen in 66.06% of cases (CDM: 17,469, GTDS: $n$ = 26,443). On average, for each patient, it was possible to map 1.39 Radiotherapy concepts, 3.94 Drug Treatments concepts, 2.02 Cancer Surgery concepts, and 1.86 Treatment Regimens concepts into the Episode table.




(**a**) (**b**)

**Figure 3.** (**a**) Sunburst plot of Disease Dynamic concepts. The segmented pie chart shows the first to fourth remission status of all patients in the Episode Domain. (**b**) Sunburst plot of Treatment concepts. The segmented pie chart shows first to fourth line therapies of all patients in the Episode Domain.

### *3.2. Vocabulary Mapping*

Furthermore, vocabulary concepts, assigned to relation types in the CDM, could be queried via the Concept_relationship table in the Standardized Vocabularies Domain. This linking of relationship types makes it possible to query additional information of a concept without this information being available in the source system. It was possible to implement 79 distinct relationships.

The "Maps to" relation was the most frequently occurring relation ($n$ = 463,880, 16.77%) (Table 2). The relation "Has priority", on the other hand, was the least represented, with five (0.0002%) events. In total, 2,765,952 oncological data-related elements could be mapped to the standard. It was possible to map 153,490 data entries to the standardized Cancer Modifier vocabulary. Among them, the concept of topography of the tumor ($n$ = 53,744, 35.01%) could be mapped most frequently.

**Table 2.** Top 10 relations with number and percentage of connected concepts.

| Relationship_ID | *n* | Percent (≈%) |
|---|---|---|
| Maps to | 463,880 | 17 |
| Mapped from | 422,273 | 15 |
| Is a | 214,233 | 8 |
| Variable to Schema | 212,210 | 8 |
| Has Answer | 160,424 | 6 |
| Has parent item | 134,531 | 5 |
| Has start date | 134,531 | 5 |
| Subsumes | 105,450 | 4 |
| Has method | 87,313 | 3 |
| Concept same_as from | 58,737 | 2 |

The OncoRegimenFinder algorithm extracted 16,303 Treatment Regimens from the documented ATC data in the source system. There were 26,443 documented protocols, similar to Treatment Regimens of the Episode Domain, in the source system. Therefore, the algorithm extracted 38.35% fewer Treatment Regimens than were stored in the source system. A brief comparison of time intervals showed that only 42% of the detected Treatment Regimens had at least one correct start or end date, assuming that the documented protocols of the source system represented the actual values. From these Treatment Regimens, 60% ($n$ = 9800) could be assigned to the regimen class Chemotherapy. The regimen class Immunosuppressive Therapy had the lowest number of events, with only 3% ($n$ = 485). The achieved vocabulary mapping score between the CDM and source system depending on the Domain can be seen in Table 3.

**Table 3.** Mapped Vocabulary by Domain.

| Domain | Vocabulary | Concept Class | Distinct Relations | *n* (With Linked Concepts) | *n* (Without Linked Concepts) | *n* (Source System: GTDS) | Mapping (%) | Mapping (%)/Domain |
|---|---|---|---|---|---|---|---|---|
| Condition | ICDO3 | ICDO Condition | Maps to, Mapped from, Is a, ICDO to Schema, ICDO to Proc Schema, Has variant, Has Topography ICDO, Has Histology ICDO, Has finding Site, Has asso morph, Concept replaces, Concept replaced by | 210,354 | 28,322 | 28,541 | 99.2 | 84.2 |
| | SNOMED | Clinical Finding | Is a, Mapped from, Maps to | 13,074 | 5419 | 7828 | 69.2 | |
| Measurement | NAACR | NAACCR Variable | Has Answer, Has parent item, Has start date, Variable to Schema, Maps to, Mapped from | 807,186 | 147,145 | 219,660 | 67.0 | 71.1 |
| | Cancer Modifier | Dimension | Maps to, Mapped from | 1694 | 865 | 1247 | 69.4 | |
| | | Metastasis | Maps to, Mapped from | 26,522 | 13,861 | 14,273 | 97.1 | |
| | | Nodes | Maps to, Mapped from | 3958 | 2033 | 7208 | 28.2 | |
| | | Staging/Grading | Maps to, Mapped from | 67,572 | 35,414 | 56,679 | 62.5 | |
| | | Topography | Maps to, Mapped from | 53,744 | 27,595 | 29,856 | 92.4 | |
| | SNOMED-CT | Staging/Scales | Is a, Mapped from, Subsumes, Maps to | 3828 | 991 | 991 | 100 | |
| | | Procedure | Maps to, Mapped from, Has component, Value mapped from, Has method, Is a | 31,154 | 5941 | 6140 | 96.8 | |
| | | Observable Entity | Is a, Subsumes, Maps to, Mapped from | 334 | 1247 | 1235 | 26.8 | |
| Observation | NAACCR | NAACCR Variable | Variable to Schema, Mapped from, Maps to, Parent item of, Has Answer | 207,144 | 85,614 | 86,442 | 99.0 | 63.0 |
| | SNOMED-CT | Morph Abnormality | Asso morph of, Maps to, Mapped from, Subsumes, Concept same_as from, Concept replaces, Is a, Concept poss_eq from | 154,358 | 30,444 | 32,022 | 95.1 | |
| | | Clinical Finding | Maps to, Has interprets, Mapped from, Has interpretation, Is a | 54,523 | 24,748 | 80,320 | 30.8 | |
| Drug | ATC | ATC 5th | Drug class of drug, Is a, Maps to, ATC—RxNorm pr lat, ATC—SNOMED eq, ATC—RxNorm | 140,182 | 26,972 | 29,205 | 92.4 | 92.4 |

**Table 3.** *Cont.*

| Domain | Vocabulary | Concept Class | Distinct Relations | *n* (With Linked Concepts) | *n* (Without Linked Concepts) | *n* (Source System: GTDS) | Mapping (%) | Mapping (%)/Domain |
|---|---|---|---|---|---|---|---|---|
| Procedure | SNOMED | Procedure | Concept replaces, Due to of, Occurs before, Has indir proc site, Maps to, Follows, Mapped frim, Value mapped from, Has surgical appr, Has access, Has temp finding, Interprets of, Has method, Has revision status, Has dir device, Has proc morph, Concept poss_eq from, Asso proc of, Focus of, Has dir porph, Has dir subst, Has proc site, Asso with finding, Using device, Has indir morph, Has complication, Has proc device, Using subst, Has intent, Has priority, Concept was_a from, Has focus, Using acc device, Subsumes, Is a, Has dir proc site, Using energy, Ha route of admin, Specimen proc of, Comoncept same_as from, Has property | 851,813 | 116 458 | 118,039 | 98.7 | 98.7 |
| Episode | HemOnc | Regimen | Is a, Mapped from, Has antineopl Rx, Has modality, Maps to, Has accepted use, Has antineoplastic, Has context, Is historical in, Has supportive med, Is current in, Concept replaces, Has support med Rx, Has local therapy, Has immunosuppr Rx, Has local therap Rx, Has immunosuppressor | 137,512 | 11,474 | 25,714 | 44.6 | 44.6 |
| Total | | | | 2,765,942 | 545,784 | 739,204 | 73.8 | 73.8 |

### *3.3. Linkage between Episodes and Underlying Clinical Events*

By using the Episode_Event table, it is possible to link the underlying clinical events to the derived Episodes. Clinical events were assigned to an Episode by their time interval. Only those events were assigned to an Episode whose examination date fell within the time interval of an Episode. In total, 2,056,721 events could be assigned to an Episode, with most of them to the Measurement Domain (Table 4). On average, 8.23 clinical events from the Measurement Domain could be assigned per Episode.

**Table 4.** Events per Episode and total events stored in the Episode_Event table.

| Domain | Events per Episode | Total Events per Episode |
|---|---|---|
| Measurement | 8.23 | 820,013 |
| Procedure | 3.29 | 477,966 |
| Observation | 4.79 | 399,128 |
| Drug | 2.94 | 238,664 |
| Condition | 1.16 | 121,044 |

Most linked events were obtained in the Measurement Domain (per Episode: 8.23, $n$ = 820,013). In total, 2,056,815 events could be assigned to an Episode.

### *3.4. CDM Application and Comparison*

Regarding the applicability of the CDM, it was tested to identify if the results of overall survival analyses across the source system and CDM were similar. The calculated survival probabilities did not differ significantly from each other ($p$-value = 0.82) and thus the H0 hypothesis was accepted. Accordingly, the median survival of a patient with breast cancer was 164 months in CDM; the calculated median survival in the source system was two months shorter (Table 5). The percentage deviation in cohort size was 1.5%, with a larger cohort included in the source system. The Number at Risk distribution can be seen in Figure 4.

**Table 5.** Descriptive statistics of overall survival of patients diagnosed with breast cancer.

| System | N | Events | Median | Standard Error | 0.95 Lower CL | 0.95 Upper CL |
|---|---|---|---|---|---|---|
| CDM | 3588 | 784 | 164.2 | 0.02 | 155.0 | 175.9 |
| GTDS | 3644 | 806 | 162.6 | 0.02 | 155.0 | 172.1 |

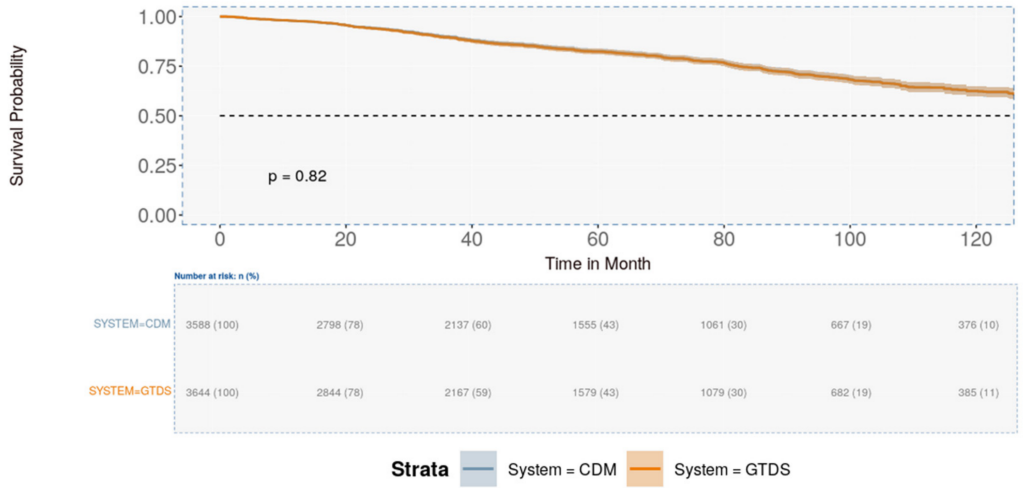



**Figure 4.** Comparison of overall survival of a breast cancer cohort between CDM and source system.

## 4. Discussion

The newly integrated Episode and Episode_Event tables, which were introduced in the last OMOP release, represent an enrichment for the representation of long-term chronic diseases, including cancer. Due to the complexity of the disease and treatment approaches, episodic modeling of the respective disease phase is useful. It is now possible to represent the course of disease according to its timeline regarding treatment and disease development. In addition, low-level clinical events of a patient can now be assigned to an Episode via their temporal reference. The use of this new data structure within the CDM increases the capabilities of oncology data analysis and visualization.

Regarding overall survival, it can be concluded that the results did not differ significantly between the systems. Nevertheless, only overall survival was evaluated in this work. The median difference in progression-free and disease-free survival was not investigated, which should be completed in the future. In addition, it must be considered that in the context of mapping to the standardized vocabularies, data entries from the source system occasionally could not be mapped to the respective standard, resulting in a reduced number of patients included in the CDM.

Additionally, it must be noted that the extension of the CDM is not yet integrated in all applications of the OHDSI community, and is only implemented on the CDM database level. Thus, the extension is also not yet integrated in the analytical and methodological toolchain provided by the OHDSI community, such as Hades, Atlas, and others, which currently limits the evaluation options of the new release. However, it can be assumed that this will be mitigated by the next major release.

In addition to implementing the Episodes, this project also addressed vocabulary mapping. Existing terminologies from the source system can be mapped to the OMOP standard. Alternatively, the terminologies in the source system can serve as a starting point for an elaborate mapping process to a completely new vocabulary, as in the case of the HemOnc mapping process. Mapping the HemOnc terminology to the OMOP standard improves the Treatment Regimen evaluation [23]. However, the algorithm developed by the Oncology Working Group (OncoRegimenFinder), which allows the translation of ATC substances to HemOnc terminology, is considered an even greater improvement. It is now possible to derive Treatment Regimens from ATC terminology, which is used in almost all European hospital systems for drug coding. In addition, by mapping source data to the OMOP–ATC and the OMOP–HemOnc standards, it is possible to revert from both terminologies to the respective RxNorm standard, which is a clinical drug dictionary for all drugs that are approved for the pharmaceutical market of the USA [20]. Through the standardized vocabularies maintained by the OHDSI 'CDM and THEMIS Working Group', it is now feasible to translate data elements from one of these three terminologies to each other, enabling international observational cancer research [23]. Furthermore, besides the ontological integration of Treatment Regimens into the CDM, the HemOnc Wiki also includes information on phase I–III clinical trials. In future releases, it is planned to include this information in the Standardized Vocabulary Metadata Repository, which would allow inference from the Treatment Regimen to the performed studies before approval [22,23]. Overall, the ontological structure of the CDM simplifies the complexity and the effort of the generation and phenotyping of cohorts; high level terms of CDM concepts can be incorporated into the query as a parameter, rather than each parameter individually, as is common in the source system.

As part of this work, source data were mapped to the Cancer Modifier vocabulary, an OMOP standard composed of different standards such as NAACCR, WHO, and SEER. However, this vocabulary is currently integrated into the CDM ontology via only two relations, which severely limits its query and analysis options, especially regarding other terminologies included in the Standardized Vocabulary Metadata Repository. Nevertheless, it can be assumed that the number of relations will increase with new vocabulary releases. Project-related mapping to the NAACCR vocabulary was challenging, since its data structure is semantically very heterogeneous compared to the source system. Therefore, only a

few data elements from the nationwide basic dataset for standardized cancer registration in Germany (ADT/GEKID), which is implemented in the source system, were mapped to the NAACCR standard. Not many terminologies that are used as standard in Europe or Germany are part of the latest vocabulary release. This complicates the mapping process, since the data must be prepared in a complex manner before they meet NAACCR vocabulary standards, leading to a loss of data during the preparation process. Therefore, in a next step, it would make sense to include other terminologies, especially those frequently used in the European or German area, such as the basic dataset ADT/GEKID, into the Standardized Vocabulary Metadata Repository of the CDM. Additionally, this would offer a general applicability of German cancer registries for data harmonization. Especially with regard to cross-cancer registry analyses, i.e., federated learning environments, this should be completed as a next step.

*Limitations*

In the context of this project, cancer-related patient data including diagnostic characteristics and prior therapies were mapped to the OMOP CDM v5.4. It has to be noted that data protection regulations in Germany make data harmonization difficult. Specifically, it is not possible to link a patient's medical record with their cancer diagnosis and map possible interactions between the development of cancer and medical records without considerable formal effort. For example, German data protection regulations impede the HL-7 import from the HIS via the GTDS interface, or the ETL into the data warehouse of the KKR of the UKE. Especially regarding personalized and predictive medicine, the problem of challenging data protection regulations should be revised. Additionally, the harmonization of EHR, EMR and registry data should be further advanced. Finally, in this study, record linkage was not considered during the mapping process. This could lead to patient duplications in the future in joint research projects with other institutions. Furthermore, concerning the calculation of overall survival, it is noteworthy that only those patients were included in the analysis whose diagnosis data could be mapped to the respective ICDO-3 standard in the Condition Domain of the CDM. Conversely, this means that all patients were excluded from the source system who could not be mapped to an ICDO-3 standard within the framework of the CDM mapping.

## 5. Conclusions

The new module, which was officially introduced by the OHDSI community in the OMOP CDM v5.4 release, is a great addition to the field of joint observational cancer research. It is currently the only CDM in the field of clinical research that includes a comprehensive standardized terminology of cancer representation and allows time-dependent episode-based modeling of disease progression. In addition, by mapping to the OMOP ontology, the source data can be enriched with additional information. This increases its application and evaluation possibilities. Furthermore, unified semantics offers the easy implementation of an AI-federated algorithm pipeline.

Nevertheless, many terminologies were included in the Standardized Vocabulary Metadata Repository that are rarely used or not used at all in the European or German areas, limiting the mapping success. This gap should be closed in the coming years to guarantee the mappability of different oncological data sources to the CDM. Especially, the inclusion of the basic oncology dataset (ADT/GEKID) in the standardized vocabulary would considerably facilitate and expand data harmonization between German cancer registries and enable joint analyses.

Additionally, it should be considered that duplicate patients can also occur in distributed research networks. These can only be clearly identified via record linkage. Therefore, future research should especially consider how to establish record linkage within the CDM across distributed research networks without contradicting the country-specific data protection regulations in place, potentially through the use of superior pseudonyms, and prepare the essential steps to enable precision medicine and precise oncology.

**Author Contributions:** Conceptualization, S.B. and J.C.; methodology, S.B. and J.C.; software, S.B.; validation, S.B.; formal analysis, J.C.; resources, S.B., F.Ü.; data curation S.B. and J.C.; writing—original draft preparation, J.C.; writing—review and editing, S.N., S.B., F.Ü., C.S.; visualization, J.C.; supervision, S.B.; project administration, C.S. and S.B.; funding acquisition, C.S. and S.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Broad consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data sharing not applicable.

**Acknowledgments:** Special acknowledgment to the Clinical Cancer Registry team. Without the administrative and technical help of the whole team this study would not have been feasible.

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

## References

1. Malone, E.R.; Oliva, M.; Sabatini, P.J.B.; Stockley, T.L.; Siu, L.L. Molecular profiling for precision cancer therapies. *Genome Med.* **2020**, *12*, 8. [CrossRef] [PubMed]
2. Liu, S.; Kurzrock, R. Toxicity of targeted therapy: Implications for response and impact of genetic polymorphisms. *Cancer Treat. Rev.* **2014**, *40*, 883–891. [CrossRef] [PubMed]
3. Liu, X.; Gao, K.; Liu, B.; Pan, C.; Liang, K.; Yan, L.; Ma, J.; He, F.; Zhang, S.; Pan, S.; et al. Advances in Deep Learning-Based Medical Image Analysis. *Health Data Sci.* **2021**, *2021*, 1–14. [CrossRef]
4. Hawkins, S.; Wang, H.; Liu, Y.; Garcia, A.; Stringfield, O.; Krewer, H.; Li, Q.; Cherezov, D.; Gatenby, R.A.; Balagurunathan, Y.; et al. Predicting Malignant Nodules from Screening CT Scans. *J. Thorac. Oncol.* **2016**, *11*, 2120–2128. [CrossRef] [PubMed]
5. Nagpal, K.; Foote, D.; Liu, Y.; Chen, P.-H.C.; Wulczyn, E.; Tan, F.; Olson, N.; Smith, J.L.; Mohtashamian, A.; Wren, J.H.; et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit. Med.* **2019**, *2*, 48. [CrossRef] [PubMed]
6. Shimizu, H.; Nakayama, K.I. Artificial intelligence in oncology. *Cancer Sci.* **2020**, *111*, 1452–1460. [CrossRef] [PubMed]
7. Eggermont, A.M.M.; Apolone, G.; Baumann, M.; Caldas, C.; Celis, J.E.; de Lorenzo, F.; Ernberg, I.; Ringborg, U.; Rowell, J.; Tabernero, J.; et al. Cancer Core Europe: A translational research infrastructure for a European mission on cancer. *Mol. Oncol.* **2019**, *13*, 521–527. [CrossRef] [PubMed]
8. Lablans, M.; Schmidt, E.E.; Ückert, F. An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium. *JCO Clin. Cancer Inform.* **2018**, *2*, 1–8. [CrossRef] [PubMed]
9. Fleurence, R.L.; Curtis, L.H.; Califf, R.M.; Platt, R.; Selby, J.V.; Brown, J.S. Launching PCORnet, a national patient-centered clinical research network. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 578–582. [CrossRef] [PubMed]
10. 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]
11. OHDSI. The Book of OHDSI; Observational Health Data Sciences and Informatics. 2021. Available online: https://ohdsi.github.io/TheBookOfOhdsi/ (accessed on 15 January 2022).
12. 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] [PubMed]
13. 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]
14. Therasse, P.; Arbuck, S.G.; Eisenhauer, E.A.; Wanders, J.; Kaplan, R.S.; Rubinstein, L.; Verweij, J.; Van Glabbeke, M.; van Oosterom, A.T.; Christian, M.C. New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* **2000**, *92*, 205–216. [CrossRef] [PubMed]
15. Altmann, U.; Katz, F.R.; Tafazzoli, A.G.; Haeberlin, V.; Dudeck, J. GTDS–a tool for tumor registries to support shared patient care. In Proceedings of the Proc AMIA Annu Fall Symp, Washington, DC, USA, 30 October 1996; pp. 512–516.
16. OncoRegimenFinder. Available online: https://github.com/OHDSI/OncologyWG/tree/master/OncoRegimenFinder (accessed on 15 January 2022).
17. HemOnc. Available online: https://hemonc.org/wiki/Main_Page (accessed on 15 January 2022).
18. Fritz, A.; Percy, C.; Jack, A.; Shanmugaratnam, K.; Sobin, L.; Parkin, D.M.; Whelan, S. *International Classification of Diseases for Oncology*; World Health Organization: Geneva, Switzerland, 2000.
19. NAACCR. Standards for Cancer Registries Volume II. *Data Standards and Data Dictionary.* Available online: http://datadictionary.naaccr.org/default.aspx?c=1&Version=21 (accessed on 24 February 2022).
20. NIH. National Library of Medicine. In *Mission*. Available online: https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-library-medicine-nlm (accessed on 24 February 2022).

21. Stearns, M.; Price, C.; Spackman, K.; Wang, A.Y. SNOMED clinical terms: Overview of the development process and project status. In Proceedings of the AMIA Annual Symposium, Washington, DC, USA, 3–7 November 2001; pp. 662–666.
22. Warner, J.L.; Cowan, A.J.; Hall, A.C.; Yang, P.C. HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals. *J. Oncol. Pract.* **2015**, *11*, e336–e350. [CrossRef] [PubMed]
23. 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]