Hubertus Wald Tumorzentrum
Universitäres Cancer Center Hamburg

HAW
HAMBURG

Hamburg University of Applied Science

Faculty Life Sciences

Study Program Health Sciences

Master Thesis

Transforming cancer registry data into an ontology-driven common data model

Evaluation of the implementation and integrability of the Observational Medical Outcomes Partnership Model in Clinical Cancer Registry

Submitted by:

Jasmin Carus

[redacted]

Hamburg, 07.04.2021

1st Examiner:

Prof. Dr. Boris Tolg

2nd Examiner:

Prof. Dr. rer. nat. Kai Rothkamm

Abstract

Introduction: The identification and incorporation of biomarkers in the treatment of people suffering from cancer is moving from guideline-based treatment towards a personalized approach in cancer care. This individual approach in cancer medicine, requires very granular inclusion criteria to conduct a study, minimizes the number of potential study participants resulting in limited power of the study. Therefore, cross-institutional data sharing to conduct collaborative studies in cancer research is becoming increasingly important. Semantic heterogeneity regarding data representation within different research institutions complicates data exchange. A possible solution could be the embedding of individualized cancer data into a common data model, which represents knowledge with a unified semantic.

Methods: Cancer registry data from the UCCH were adapted to the standardization of the OMOP CDM v.6.0. In addition to the implementation, the CDM was tested with regard to two use cases. The first use case examines to which extent the ADT/GEKID data set can be integrated into a CDM for cross-institutional data exchange. The second use case examines if key performance indicators of the urinary bladder module within the scope of organ centre certification, with the goal to ensure quality of care in hospital, can be represented using CDM.

Results: A total of 1936055 data entries from the source system could be implemented as standardized concepts in OMOP CDM v.6.0. 80626 data entries were implemented as custom vocabularies. Through the ontologies, which are already integrated in the OMOP CDM, 78.5% of the ADT/GEKID base data set can be mapped. When determining the key performance indicators of bladder module in the context of organ centre certification, the CDM seems to slightly underestimate when including the arithmetical mean ($\bar{x}$ = -0.5).

Conclusions: The OMOP CDM v.6.0 is a suitable tool for data harmonization in operational databases with heterogeneous semantic. For the transmission of cancer data within the scope of ADT/GEKID, the model is suitable for 78.5% of the data. The transfer of the ADT/GEKID into the vocabulary of the OMOP CDM could contribute a big step in European data integration. For the determination of organ-specific key performance indicators in the context of a certification, the model is suitable. However, the vocabulary of the CDM have to be extended for this use case.

Contents

List of Figures

List of Tables

List of abbreviations

A

ADT/GEKID. *standardized data set of the working group of german tumor centres*

ATC. *Anatomical Therapeutic Chemical Drug Classification*

B

BRCA. *Breast Cancer, early-onset*

C

CAP. *College of American Pathologists*

CCC. *oncological centres of excellence*

CDISC. *Data Interchange Standards Consortium*

CDM. *Common Data Model*

CEA. *Carcinoembryonic antigene*

CML. *chronic myeloid leukemia*

CTCAE. *Common Terminology Criteria of Adverse Events*

D

ddA. *doxorubicin*

DKFZ. *German cancer research centre*

DKG. *German Cancer Society*

DNA. *Deoxyribonucleic acid*

D-rd. *dartumumab, lenalidomide and dexamenthasone*

E

EERB2. *erb-b2 receptor tyrosine kinase 2*

EHR. *electronical health records*

EMR. *electronical medical records*

ETL. *Extract, Transform, Load*

F

FDA. *US Food and Drug Administration*

G

GTDS. Giessen Tumor Documentation System

H

HER2. *epidermal growth factor 2*

HIS. *Health information system*

HKR. *Hamburg Cancer Registry*

HmbKrebsRG. *Hamburg Cancer Registry Act*

HRQoL. *Health-related quality of life*

I

ICD-10. *International Classification for Diseases and Related Health Problems, 10th revision*

ICD-O-3. *International Classification of Diseases for Oncology, 3rd Edition*

K

KFRG. *Early detection of Cancer and Cancer Registry law*

KKR. *Clinical Cancer Registry*

KPI. *Key performance indicator*

L

LUT. *Lookup Table*

M

MONTAC. *Manual of tumour nomenclature and coding*

mRNA. *messenger ribonucleic acid*

N

NAACCR. *North American Association of Central Cancer Registries*

NaKo. *National Cohort*

NCIt. *National Cancer Institute Thesaurus*

NIH. *National Institutes of Health*

NKP. *National Cancer Plan*

NLM. *National Library of Medicine*

O

OHDSI. *Observational Health Data Science and Informatics*

OMOP. *Observational medical outcomes partnership model*

OPS. *Operation and procedure key*

P

PCORnet. *Patients-Centered Clinical Research Network*

PROM. *Patient-reported outcome measures*

R

Rd. *lenalidomide and dexamethasone*

S

SDTM. *Study Data Tabulation Model*

SEER. *Surveillance, Epidemiology, and End Results Program*

SSIS. *SQL Server Integration Services*

U

UCCH. *University Cancer Centre Hamburg*

UKE. *University Hospital Hamburg*

UMLS. *Unified Medical Language System*

USA. *United States of America*

W

WHO. *World Health Organization*

X

XML. *Extensible Markup Language*

XSD. *XML Schema*

1 Introduction

Cancer rates among the most severe diseases nowadays while being the cause of death in 25% of death cases in 2019, in Germany (Statistisches Bundesamt 2021). Since there is no particular pattern of especially severe affected people, cancer poses a risk for a broad group of people. Therefore, cancer research is highly relevant in today's medical and healthcare research in order to alleviate the suffering of people with cancer as far as possible and to increase life expectancy. In recent decades, continuous progress has been made in the complex medical treatment of patients that suffer from cancer. A reason for this is the extensive fundamental research in the field of molecular genetics regarding the development of malignant tumours. The inclusion of a variety of genetic mutations in the field of cancer research allows a growing individualized approach to control or heal cancer, but this also means that recruiting of study participants within a research institution in order to conduct studies is almost impossible because the inclusion and exclusion criteria of studies are very granular. To circumvent this problem, data sharing to conduct joint studies among different research institutions is becoming increasingly important. But due to the heterogeneous representation of data in institutions, the homogeneous exchange of data as a basis for scientific analysis is challenging.

The detection of molecular biomarkers has greatly improved the understanding of cancer, the options of treatments and the early detection of cancer. Biomarkers are biological molecules found in DNA, blood, etc., from which prognostic or diagnostic statements can be derived. Consequently, in recent years, biomarkers of tumour cells have become the focus of cancer research. As a result, targeted cancer therapy as part of complex medical treatment is getting more and more important in cancer treatment. In contrast to current chemotherapies, which essentially inhibit the cell division of most cells in the human organism, targeted therapies act directly on tumour growth. However, the detection of more and more biomarkers and the derivation of new cancer therapies lead to the fact that cancer therapy transforms to a personalized treatment approach. This also means that clinics and research institutions have problems in recruiting subjects who meet the specific characteristics regarding biomarkers and therefore fit the inclusion criteria of the conducted study. Inclusion of an appropriate study population to achieve significant results, is limited by the complex inclusion and exclusion criteria of designed studies investigating the efficacy of targeted therapies within a research institution. In modern cancer research data exchange or the establishment of analysis pipelines based on a homogeneous data semantic in joint networks of individual research institutions for the investigation of biomarkers and efficacy testing of targeted therapies is indispensable. Hence, there is the need of establishing distributed large-scale oncology networks. As Eggermont et al. have noted, there is a “need for creating a uniform platform for translational cancer research to bring together enough centres to generate the critical mass of patients, expertise and resource required to make a significant breakthrough in cancer care” (Eggermont et al. 2019:. 523). Lablans, Schmidt and Ückert from *German Cancer Research Centre* (ger. Deutsches Krebsforschungszetrum = DKFZ) identified a number of challenges for the establishment of such networks. Due to the existence of different data protection laws worldwide, merging data is challenging. One possibility would be the exchange of aggregated data sets only, but this could lead to a limitation of statistical significance in clinical trials and therefore to limited interpretation of the results. Furthermore, depending on the system, there are different technical requirements (e. g. documentation system, etc.) that can make data exchange difficult. However, the greatest challenges, in terms of a global view, lie in the area of semantic heterogeneity (Lablans et al.2018: 2). Reliable clinical research requires a comprehensive and well documented clinical data management. But clinical data management primarily aims to support everyday clinical practice. Therefore, existing terminologies are often modified depending on the application scenario (ICD10-GM vs ICD10-CM vs ICD-10) or new terminologies are developed, which are not used internationally (e.g. ger. Operations- und Prozedurenschlüssel = OPS). A possible approach to bridge the semantic heterogeneity of different data representation systems could be the transfer of electronical health records (EHR), electronical medical records (EMR), or registry data to a Common Data Model (CDM). Through CDM, knowledge can be represented in a unified form. The data that flows in a CDM has to be transform in a standardized format defined by the CDM through entities, attributes and relationships. This enables a comparability of data within the CDM, despite the integration of different operational data sources or the different use of classification systems or ontologies. The long-term goal of this work is to homogenize individual cancer data from disparate operational databases due to standardized ontologies/vocabulary and extend existing terminologies in source system through mapping, with ontologies which are integrated in the CDM. An objective of this approach is that institutions which have successfully integrated an ontology-driven

CDM into their technical infrastructure can exchange/analysis data within the CDM community in a simplified way with less costs. Therefore, the more items from the source system are translated into a CDM, the more granular patient collectives can be formed and these can be exchanged with other institutions in a simplified manner for conducting cancer research. For this purpose, part of the clinical cancer registry data of University Hospital of Hamburg (ger. Universitätsklinikum Hamburg Eppendorf = UKE) will be translated to the *Observational and medical outcomes partnership model* (OMOP) CDM version 6.0 and then analyzed for utility over its actual application scenario, which represents cross-institutional clinical research. For this purpose, the CDM is being tested with regard to its applicability for determining key performance indicators (KPI) within the scope of a centre certification. And theoretically examined the extent to which OMOP CDM can be used to transmit cancer data cross-institutional using the ADT/GEKID data set. Due to the need of shared research aiming in an increased likelihood of getting statistically powerful results, the fact of very granular in- and exclusion criteria in cancer research and the different cancer data representation worldwide, the application of CDM is getting more important. In the last years, several CDMs have been developed for the transmission of EHRs and EMRs with the purpose of conducting observational studies or active surveillance programs in the field of clinical research. This paper investigates the implementation and integration of the OMOP model into the technical infrastructure of the clinical cancer registry of the *University Cancer Centre Hamburg* (UCCH). The objective of this work is to find out:

1. To what extent can cancer registry data be transferred to the OMOP CDM v.6.0 model?
2. Can the OMOP CDM v.6.0 model be integrated beyond its utility into the infrastructure of a clinical cancer registry?

2 Background

This section highlights areas of current cancer research and the work of UCCH and its importance of cancer research to the northern region of Germany. Furthermore, the *clinical cancer registry* (ger. Klinisches Krebsregister = KKR) and its responsibilities are described. The administered data of the KKR form the data basis of this paper and is the starting point of the *Extract, Transform Load* (ETL) process for the translation of the source data into the OMOP model. The OMOP model was developed by the

Observational Health Data Science and Informatics (OHDSI) collaborative with the aim of facilitating the exchange of data from medical research institutions and producing clinical evidence from them. The OMOP CDM v.6.0 model is presented in detail and its involvement in current research projects is described.

2.1 Current Cancer Research

Current cancer research focusses on molecular pathology detection of biomarkers. The term biomarker is described by the National Cancer institute as a "… biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition. Also called molecular marker and signature molecule" (National Cancer Institute 2021). Biomarkers comprise a large number of molecules. However, in the context of cancer detection, the most important biomarkers are located in areas of DNA, enzymes, mRNA, metabolites, cell surface receptors or in transcription events (Wu/Qu 2014: 2964). The detection of cancer biomarkers and associated analytical techniques (e.g. liquid biopsy) have raised continuously and helped to promote the development of different areas of cancer care in recent years. A study by Hall et al. investigated whether certain biomarkers could be associated with an increased likelihood of developing breast or ovarian cancer. It was proven that the BRCA mutation on chromosome 17q21 leads to an increased probability of developing breast or ovarian cancer. In addition, it was shown that the BRCA mutation is inherited in families (Hall et al. 1990: 1684). The study by Hall et al. shows how biomarkers can be used to evaluate the probability of developing cancer. Moreover, biomarkers can also be used to determine the prognosis of a disease. Paik et al. developed a multigene assay to predict recurrence breast cancer by patients who were treated with tamoxifen and are node-negative (Paik et al. 2004: 2818f.). In a systematic review by Locker et al., it was pointed out that active monitoring of carcinoembryonic antigene (CEA) by patients who suffer from metastatic colorectal cancer, during systemic therapy, gives good indication of treatment response or lack of response (Locker et al. 200: 5314).

The identification of biomarkers in cancer treatment, and the derivation of treatment strategies that favor the patient's own immune response or the death of cancer cells, are referred as targeted therapies. Already in year 2000, the first monoclonal antibody, trastuzumab, was approved as targeted therapy in the European Union for the treatment of metastatic breast cancer in patients overexpressing EERB2. Researchers found out that administration of a monoclonal antibody, such as trastuzumab, which binds to epidermal growth factor 2 (HER2), in combination with chemotherapy leads to improved survival of patients with metastatic breast cancer (Slamon et al 2001: 784f.). At the same time, other targeted therapies, such as the drug imatinib, have been developed. Patients with a chromosomal translocation in chromosome 22 develop the fusion protein BCR-ABL, leading to the uncontrolled proliferation of white blood cells, which promotes the development of chronic myeloid leukemia (CML). Researchers were able to prove that the tyrosine kinase inhibitor imatinib resulted in a hamaetological (77% of studycohort) and cytogenetic response (53% of studycohort) in patients with CML who were treated with imatinib (Druker et al. 2001: 1034).

The molecular prognostic determination of biomarkers and its derivation of targeted therapies are associated with an increased likelihood of response in combating the tumour condition. This approach is referred as personalized cancer medicine. The term personalized does not refer to the person with the disease but describes the stratification of certain patient collectives depending on their genetic constellation. From the patient collectives that share the same genetic variations, the prognostic factor can be determined, from which the possibility of targeted therapies can be derived, associated with an increased probability of response in the treatment of cancer (Damm 2011: 7f.). To form sufficiently large patient collectives that share the same genetic variations, homogeneous semantic is essential in terms of cancer data representation. As more and more biomarkers are identified, patient collectives are becoming more granular and cross-institutional data sharing for research purposes is becoming imperative to derive clinical evidence. A possible approach for homogenous data representation in cancer research is the OMOP model, which assigns a standardized value to each data entry, which is related to other standardized values by relationship types. This ontology-based vocabulary provides cross-institutional analysis capabilities based on the CDM as well as simplified data exchange due to standardization. CDMs thus, represent a powerful tool to solve the problem of heterogeneous semantic in current cancer research.

2.2 University Cancer Centre Hamburg

Considering increasing trends in incidence of cancer and an increase of life expectancy under cancer treatment, cancer research and cancer registries tend to be the key policies to evaluate the impact of therapy strategies and monitor actual trends in cancer development in Germany. Therefore, the Ministry of Health in Germany developed the

*National Cancer Plan (*ger. Nationaler Krebsplan = NKP*)* in 2009 for further improvements in medical supply for humans who are suffering from cancer. NKP promoted the establishment of a competence network of some university hospitals in Germany and formed the base for the *Early detection of Cancer and Cancer registry law (*ger. Krebsfrüherkennungs- und registergesetz = KFRG*)* (Bundesgesundheitsministerium, 2019). KFRG became effective in 2013, nationwide. Since 2014 the *Hamburg Cancer Registry Act (*ger = Hamburgisches Krebsregistergesetz = HmbKrebsRG) regulates data exchange between health care providers and the *Hamburg Cancer Registry (*ger. Hamburgisches Krebsregister = HKR*)*. On the base of a standardized dataset (ADT/GEKID) hospitals and health care providers (e.g. physicians) have to report about their treated patients concerning course of disease, therapy strategies and other key facts about patient’s cancer diagnosis and therapy (Arbeitsgemeinschaft Deutscher Tumorzentren 2020).

The establishment of a competence network within the framework of the NKP was achieved in year 2009 on initiative of the *German cancer society* (ger. Deutsche Krebsgesellschaft = DKG). The association of several oncological centres of excellence in a network (ger. Onkologische Spitzenzentren = CCC) is intended to increase the conceptual cooperation, develop new treatment standards, create a compatible documentation system infrastructure, upgrade biobanks, increase the promotion of cancer research and the expansion of translational cooperation. Furthermore, it is pursued to increase the active transfer of collected knowledge within the network to the vicinity of the CCC to ensure adequate patient care in the region (Netzwerk Onkologische Spitzenzentren 2020). As one of the 13 members of the competence network, the UCCH covers the provision of new medical approaches in cancer therapy and cancer research in the northern region of Germany. Since its foundation in 2007, UCCH represents the organizational unit regarding cancer at the UKE. It is responsible for the assessment and recommendation of cancer treatment through the development of guiding principles and research in this field. Furthermore, it develops cooperation with other partners, with the aim of sharing information and achieving a common structure for preclinical research and patient’s wellbeing in the region (University Cancer Centre, 2019). The overall goal is the common share of these findings within these network aiming in an increased patient care and treatment response (Deutsche Krebshilfe 2019).

UCCH research activities

To achieve this, in addition to the treatment of tumour diseases, clinical cancer research is crucial. UCCHs research activities focuses on dissemination and metastasis of malignant neoplasms; primarily entities such as leukaemia and lymphomas, prostate cancer, gastrointestinal tumours, head and neck tumours or neuro-oncological tumours are areas of interest within the scope of clinical cancer research. Furthermore, UCCH pursued to incorporate clinical research results into translational approaches and to link basic research with the latest medical knowledge and generate new treatment guidelines from this. In many studies in which UCCH has been involved, patient-reported outcome measures (PROM) have taken place to uncover potential differences between patients' and physicians' view and to cope these differences in clinical practice aiming in an increased wellbeing of patients. Thus, UCCH examines which treatment paths with regard to a tumour disease increase quality of life of patients. Therefore, a multicentre study was conducted to investigate the extent to which the therapy regimens daratumumab, lenalidomide, and dexamethasone (D-Rd) compared to treatment with lenalidomide and dexamethasone (Rd), have an impact on health-related quality of life (HRQoL) in patients who suffer from multiple myeloma and are not transplantable. It was found that subjects who received the treatment regimen of D-Rd had an over constant higher general health score. In addition, subjectively perceived pain was lower in the D-Rd arm compared to the group of subjects treated with Rd (Perrot et al. 2021: 228, 230). Incorporating PROM tools, such as HRQoL, into clinical research is an important element to broaden healthcare by patient perspective (Staniszewska, et al. 2012: 80). In addition to the application of PROM techniques in clinical research, UCCH investigated the connection of molecular genetic tumour detection in multiple studies and how these findings could be embedded in clinical treatment strategies. For instance, a multicentred study by Murthy et. al found out that patients with metastatic breast cancer whose cancer cells had a positive HER2 factor and were found to be progressing after first-line therapy, had good response rates in terms of progression-free survival when treated with tucatinib, trastuzumab, and capecetabine compared with a placebo, trastuzumab, and capecetabine (Murthy et al. 2019: 600ff.). Efficacy studies to evaluate treatment success are an essential component of UCCHs clinical research. Next to efficacy studies, UCCH is also part of cancer epidemiology projects. These research projects offer the potential to identify risk factors in the population and actively incorporate these findings into cancer prevention programs. UCCH is part of the *National Cohort* (NaKo), a long-term study that includes patient surveys and medical examinations and aims to a better understanding of diseases such as diabetes, cancer or cardiovascular diseases that have a high prevalence and a high incidence in the population. NaKo is trying to transfer these findings into clinical practice (German National Cohort Consortium 2014: 371f.). However, care research in the clinical context of patients suffering from cancer is also a big concern at UCCH. A study by Mehnert et al. indicated that psycho-oncological care in patients who are suffering from cancer may be an important component in decreasing the prevalence of mental illness in this patient cohort. This study estimated the 4-week prevalence of the most common mental disorders in patients who had cancer. It was found that 31.8% of the study participants suffered from a mental disorder, with anxiety disorder being the most common among the participants. Whereby the prevalence varies in relation to the respective entity. Patients with breast cancer had a higher prevalence of mental disorders compared to patients with a pancreatic tumour (Mehnert et al. 2014: 3542ff.). Such findings are essential within the context of the complex treatment of a tumour disease. Pharmaceutical studies, efficacy research, epidemiological approaches and surveillance and monitoring strategies with the aim of an increased treatment success and patient's wellbeing build the fundamental approach of the UCCH. Studies in UCCH are often conducted in association with other research institutions. Data representation in a uniform format and the exchange of data for study purposes reduces costs and duration of the study period. Therefore, UCCH tries to link their data on embedding semantic knowledge representations (e.g. UMLS – Unified Medical Language System) systems.

Clinical Cancer Registry

To evaluate treatment success, the structured processing of cancer data is essential. Since 2014 the UCCH has been furthermore obliged to report all cancer-related treatments and diagnoses to the HKR as part of the HmbKrebsRG. To ensure the fulfillment of NKP, the UCCH founded the department *Clinical Cancer Registry of the UKE* (ger. Klinisches Krebsregister = KKR). It deals with the standardized documentation of all cancer patients at the UKE. Moreover, these cancer data are the base for further responsibilities of the KKR. Primary responsibilities are:

I. **Reporting to the HKR**

Reporting of cancer cases is statutory obligated with the purpose of monitoring cancer occurrence in the population but also to measure changes of treatment success or failure. HmbKrebsRG expects reporting within a time frame of 8 weeks regarding defined notification types (new diagnosis of cancer, treatment start/end, status change, death).

II. **Quality assurance**

UCCH assures quality regarding treatment of cancer. This is achieved due to a standardized operation principle within the documentation process but also by certification of medical centres which treat cancer patients. KKR provides data and analyzes them for certification purposes. Additionally, it develops indicators for measuring quality assurance in treatment of cancer patients and reports it within the institute.

III. **Research**

KKR provides data for research purposes to physicians or scientific employees, or conduct analyzes for them. In addition, KKR annually reports regarding quality indicators and the development of cancer burden in UKE.

2.2 Observational Health Data Sciences and Informatics (OHDSI) collaborative

One possible solution for the uniform representation of medical data is offered by CDMs. They represent knowledge due to a common language through integration of standard concepts, entities and specifications. Many CDMs come along with analytical applications. Thus, the integration of heterogenous operational databases into a CDM enables the use of developed analytical applications for the CDM (see Fig. 1). The integration of CDM and its applications into a medical data warehouse reduces time and costs for the company by expanding individual analytical tools and software by applications of the CDM. In a study by Garza et al. the most common CDMs (Sentinel Common data model v.5.0, Patient-Centred Clinical Research Network (PCORnet) Common Data Model v.3.0, Observational medical outcomes partnership model (OMOP) v.3.0, Data Interchange Standards Consortium (CDISC), Study Data Tabulation Model (SDTM v.1.4.)) in the clinical research domain were evaluated in terms of completeness, integrity, flexibility, integrability, and implementability for EHR-based longitudinal registry data. It was found that the OMOP CDM v.3.0 achieved the best scores with regard to the evaluation criteria (Garza et al. 2016: 334f., 340). Based on the evaluation of Garza et al., the OMOP CDM v.6.0 is used in this paper. Figure 1 shows the functionality of a CDM using the OMOP CDM as an example. It illustrates the goal of this work to achieve homogenization of data from heterogeneous operational databases (EHR, EMR, Registry data) by using standardized ontologies. OMOP CDM can be applied to 1. use existing ontologies and mapping of the CDM to integrate ontologies in one's data pool that are not generated by the operational data sources themselves and 2. to use the existing analysis tools of the OHDSI collaborative.

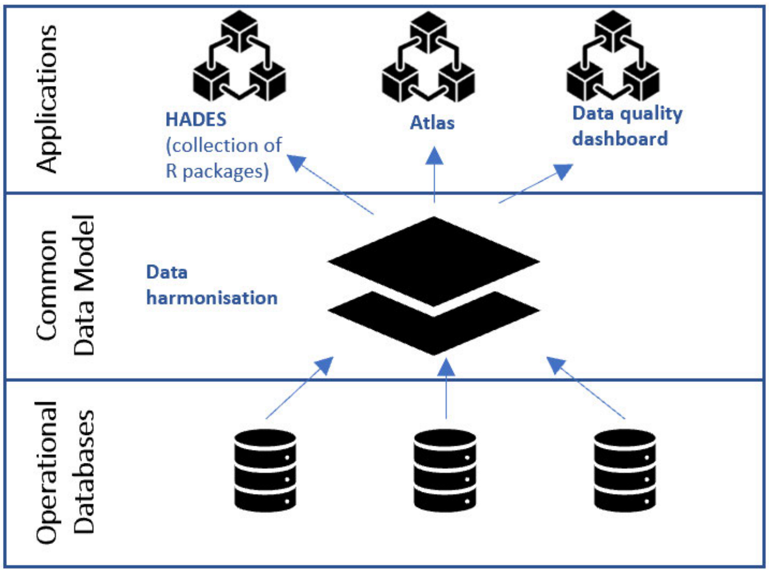
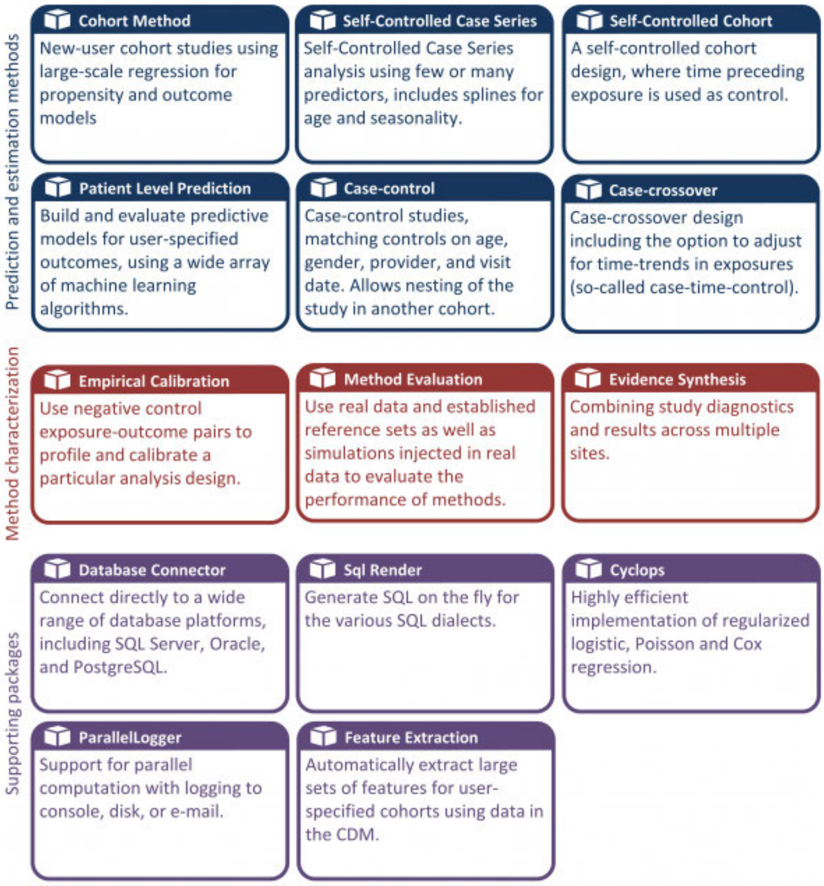


Figure 1: Functionality of a CDM using OMOP as an example

The OMOP model was developed by OHDSI. It is a multi-stakeholder interdisciplinary collaboration founded in 2014. It arose from the public-private partnership of the US Food and Drug Administration (FDA). Within the framework of this private partnership, a data model should be created, for which it is possible to conduct medical product safety surveillance with observational healthcare data across institutes. After the end of the FDA funding, it was decided to develop a collaboration (OHDSI) and adopt the CDM as an open-source project trying to integrate the CDM in scientific applications. Meanwhile, this collaboration consists of an international network of researchers and over 100 disparate observational health databases from 19 different countries. Due to the lack of unique EHR and EMR and the absence of consistent patient-level data in observational databases, this cooperation develops technical solutions for the representation of uniform medical data from different source systems (OHDSI 2020: 3f.). It provides open-source applications with the goal of strengthening the community’s self-determination and derive evidence to be considered in clinical questions (OHDSI, 2019). Observational databases vary regarding different clinical demands, purposes and design issues. Due to various approaches, the use of EMR and EHR transformation into a CDM for reliable and consistent analyses and the common share and comparison of these findings for further improvements in clinical research is essential. In addition, harmonization into a CDM declines capture bias and large numbers of observed patients in a study indicates a higher statistical power of the study. OMOP CDM v.6.0 offers the identification of a population cohort with a predefined similarity but also the characterization of these cohorts regarding specific parameters (biomarkers, entity, stage group, histology, etc.). Furthermore, analytical tools developed by ODHSI, which build upon this CDM provide advanced analytics in prediction of (measured) outcomes in individual patients and estimate the effect size of these intervention (see Fig. 2). With these application possibilities OMOP CDM v.6.0 offers a powerful data model for clinical research and the handling of clinical data (OHDSI 2019).



Cohort Method
New-user cohort studies using large-scale regression for propensity and outcome models

Self-Controlled Case Series
Self-Controlled Case Series analysis using few or many predictors, includes splines for age and seasonality.

Self-Controlled Cohort
A self-controlled cohort design, where time preceding exposure is used as control.

Patient Level Prediction
Build and evaluate predictive models for user-specified outcomes, using a wide array of machine learning algorithms.

Case-control
Case-control studies, matching controls on age, gender, provider, and visit date. Allows nesting of the study in another cohort.

Case-crossover
Case-crossover design including the option to adjust for time-trends in exposures (so-called case-time-control).

Empirical Calibration
Use negative control exposure-outcome pairs to profile and calibrate a particular analysis design.

Method Evaluation
Use real data and established reference sets as well as simulations injected in real data to evaluate the performance of methods.

Evidence Synthesis
Combining study diagnostics and results across multiple sites.

Database Connector
Connect directly to a wide range of database platforms, including SQL Server, Oracle, and PostgreSQL.

Sql Render
Generate SQL on the fly for the various SQL dialects.

Cyclops
Highly efficient implementation of regularized logistic, Poisson and Cox regression.

ParallelLogger
Support for parallel computation with logging to console, disk, or e-mail.

Feature Extraction
Automatically extract large sets of features for user-specified cohorts using data in the CDM.

Figure 2: HADES - an open-source R packages collection (HADES 2021)

To ensure evidence-based research, unified representation of data is essential. In order to achieve this, data items that are included in the CDM must be implemented in a standardized format represented by ontologies. Often common ontologies/terminologies are adapted to country-specific requirements, such as *International Classification for Diseases and related Health Problems* (ICD-10), provided by the *World Health Organization* (WHO), which is used in the United States of America (USA) as a clinical modification (ICD-10-CM), whereas in Germany it has been adapted to the needs of the German healthcare system and is used as a German modification (ICD-10-GM) (OHDSI 2020: 55ff.). This leads to the fact that some of the data, which is included in an ontology, is represented differently by country-specific modifications. In the OMOP CDM, a uniform standardization is mandatory. Each data value can be mapped to a standard and a source concept. The standard concept is defined by the CDM developers (see chapter 3.1), whereas the source concepts are not defined in detail, but they must belong to the corresponding domain and have to be included in the vocabulary of the CDM. The vocabulary in the CDM describes a common repository of all vocabularies or ontologies that are available in a standardized form and are revised and maintained by the OMOP CDM developers. In addition, these standardized concepts are linked to each other by relation attributes, which considerably expands the analysis options of the integrated concepts in the CDM. The integrated standardized vocabulary in CDM is elementary because it enables comparability of different data sources by mapping to a unified repository that is used exclusively by the community. This standardization regarding the uniform representation of medical data, forms the basis for all further topics of the OHDSI collaboration. In addition to standardization, the quality of the data has an impact on the quality of the results. In CDM, generic data quality is divided into three components: Conformance, Completeness and Plausibility. All three components can be validated or verified. (OHDSI 2020: 292f.). To ensure this, it is possible to check the CDM for conformance, completeness and plausibility. For this purpose, ACHILLES was developed, a software that performs rule-based checks to determine whether the data quality components have been met. Next to high-level-checks, individual-level-checks during the ETL-Process are possible. Correct standardization and the application of quality checks during ETL-Process are essential for conducting further research with the OMOP model.

OMOP CDM v.6.0

The OMOP CDM v.6.0 is a patient centric model which is designed as a relational model. As Figure 3 shows, it can be divided into seven thematic areas. In this work, however, some areas of the OMOP model are not illuminated, because 1. the source data do not provide these information (e.g. standardized health economics), or 2. the filling of the tables is only intended after successful implementation (e.g. results schema in standardized derived elements area), or 3. the filling of the tables was done for the reason of storing information of the source system (e.g. standardized metadata). The ***standardized clinical data*** tables are linked to the *person* table, which allows a longitudinal view on all relevant health care events of one person. An exception is the ***standardized health system data*** which is linked to the events of various domains. Events are presented through standard and source concepts categorized into a specific domain and defined through a standardized vocabulary (e.g. SNOMED-CT). Source data is mapped to a standard concept format by incorporating standardized ontologies, which are integrated into the CDM's ***standardized vocabularies*** theme*.* Source_values in each table provide the source data in its original form. Due to the non-standard form of source_values, these records are unsuitable for advanced analytics regarding outcome measurements of patients or cohorts, but essential regarding the measurement of quality assurance. If the source values are available in a standardized form and the source vocabulary is integrated into the CDM, it is possible to map the source data as source concepts in addition to the standard vocabulary (e.g. SNOMED-CT is designated as standard in the area of condition, but it is also possible to integrate the ICD-10-GM as source concepts in condition domain) with regard to the standardization of the OMOP model. Tables in the OMOP CDM v.6.0 are considered read-only, except tables in ***Results schema***. In the *cohort* and *cohort_definition* table, individual definitions of groups of interest are possible (e.g. primary cases for certification issues) (OHDSI 2019).

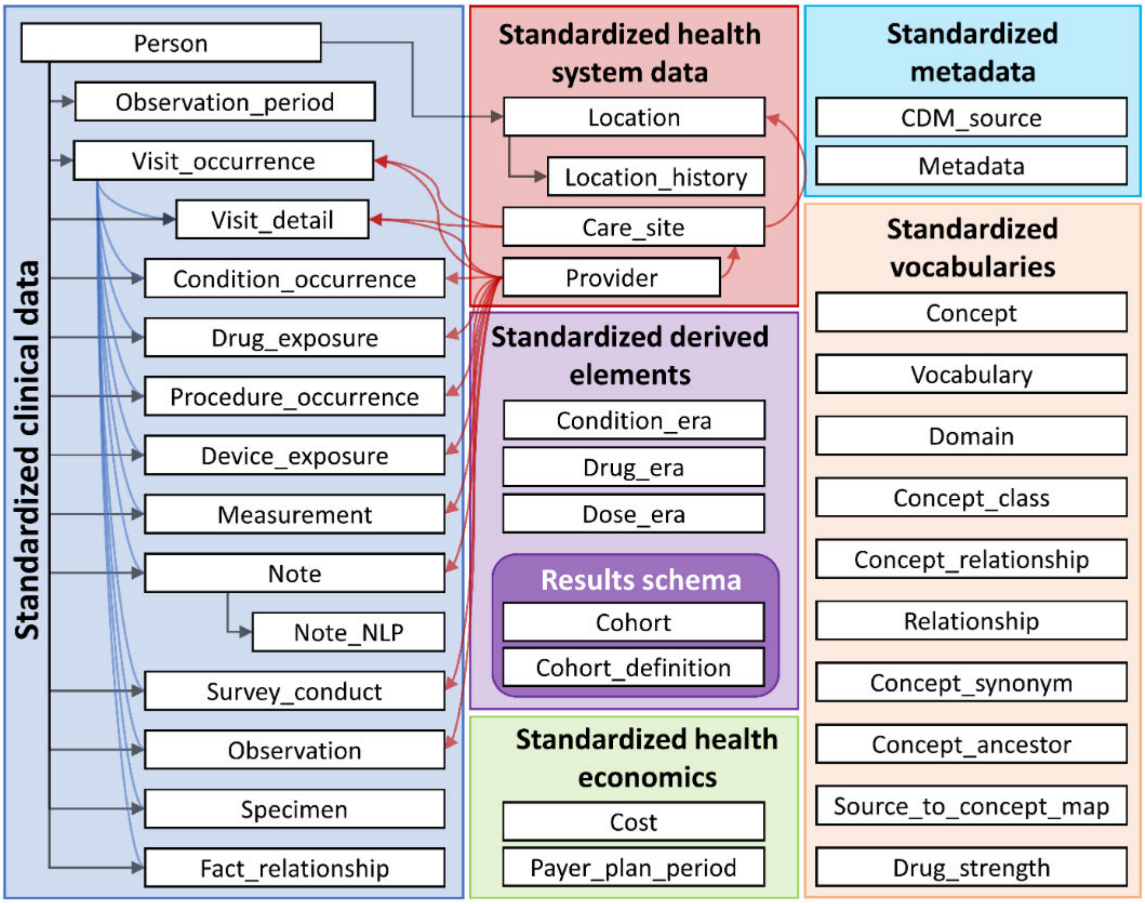


Figure 3: Relational Design of OMOP CDM v.6.0 (OHDSI 2020: 32)

Analytical tools, developed by OHDSI, on top of CDM support different use cases e.g. patients level prediction aiming at the application of machine learning algorithms for precise medical decision, clinical characterization for disease, treatment and quality improvement and the estimation of population-level-effects for surveillance purposes or comparative effectiveness.

Oncology Module Extension to OMOP CDM v.6.0

Research in cancer or cancer data representation often requires a more detailed preparation of the existing data. Thereby, the representation of cancer data can be compared with the record of a chronic disease. In both cases, episodic modeling of the disease is required. The course of a cancer disease is influenced by a constellation of various conditions. For instance, cancer diagnosis is defined through a combination of histology and topography of the tumour. Other diseases, for which OMOP CDM v.6.0 was developed, do not require this information. In addition, the treatment approach might differ due to cancer modifiers which define the course of cancer in more detail (e.g. stage, grade, tumour biomarkers). Furthermore, treatment approaches are influenced by cancer entity and administered due to defined order or cycle. To ensure comprehensive cancer data representation, an extension of OMOP CDM v.6.0 is necessary. Therefore, data abstraction is needed which is often not present in source data. The OHDSI Oncology Module at work extends OMOP CDM v.6.0 to ensure disease and treatment abstraction with the aim to support information from source data with the required granularity but also for standardized cancer related analytics (e.g. Overall survival) (OHDSI 2018: 26f.).

Through the oncology module, new ontologies have also been implemented in the CDM that closely correlate with the treatment of cancer patients (Belenkaya, et al. 2021: 13). These are:

- World Health Organisation International Classification of Diseases for Oncology, 3rd Edition (**ICD-O-3**)
- **HemOnc.org** – medical Wiktionary of intervention, regimens and information in field of hematology and oncology
- North American Association of Central Cancer Registries (**NAACCR**) – data dictionary
- Anatomical Therapeutic Chemical Drug classification (**ATC**)
- College of American Pathologists (CAP) (College of American Pathologists, 2021)
- Nebraska Medicine Clinical Ontology Application (Nebraska Lexicon)
- National Cancer Institute Thesaurus (NCIt)

The highlighted ontologies are described in detail in chapter 3.1. CAP, Nebraska Lexicon, and NCIt ontologies are not discussed in detail in this elaboration because they were not included as vocabulary in the ETL as part of the implementation of the oncology module.

Belankaya et al. extended OMOP CDM v.6.0 through an *episode* and the underlying *episode_event* table. Furthermore, cancer diagnosis is presented through pre-defined concepts (combining histology + topography = ICD-O-3) in the *condition_occurrence* and *episode* table, which belongs to the ***standardized clinical data***. Cancer treatment events (e.g. drugs) are stored in *procedure_occurrence* and *drug_exposure*, whereas disease and treatment episodes (e.g. hormone therapy) are presented in the new *episode* table. A linkage between disease, treatment event and episode and standardized clinical data tables of CDM can be done in *episode_event* table. Additional treatment characteristics or diagnostics, which are handled as modifiers in oncology module, are stored in the *measurement* table through extension of variables modifier_of_event_id and modifier_of_field_concept_id in this table (see Fig. 4) (ebd.: 13ff.).

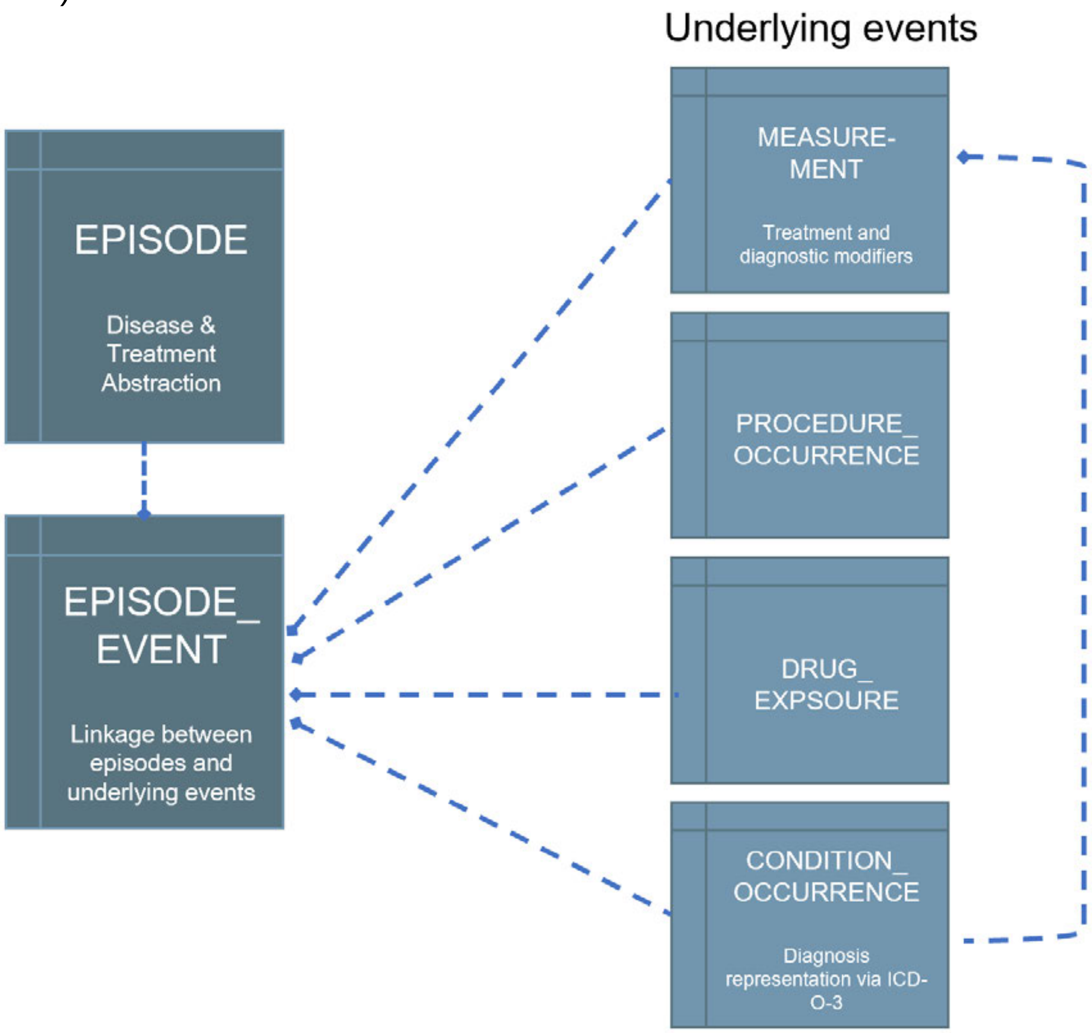

Underlying events
EPISODE
Disease & Treatment Abstraction
EPISODE_ EVENT
Linkage between episodes and underlying events
MEASURE- MENT
Treatment and diagnostic modifiers
PROCEDURE_ OCCURRENCE
DRUG_ EXPSOURE
CONDITION_ OCCURRENCE
Diagnosis representation via ICD-O-3

Figure 4: Linkage between oncology module and underlying events of CDM (OHDSI 2018)

Application of OMOP CDM

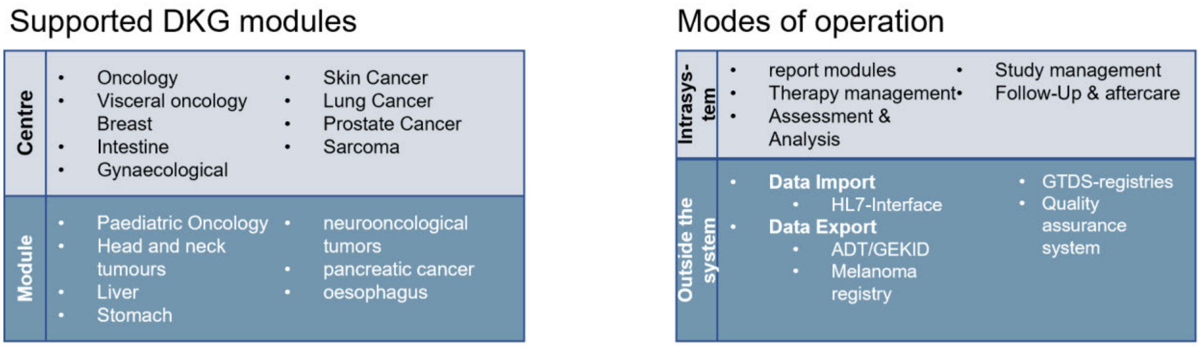
After successful implementation of ETL in technical infrastructure of the company, software applications, which build upon CDM, offer powerful tools in the field of observational research. In a study by Voss et al. the feasibility and application possibilities of CDM were investigated for different observational databases. It was found that, despite minor data loss during data mapping, the effectiveness of cohort formation is significantly accelerated by the implementation of a CDM. In addition, the barriers to a common systematic approach in observational research in different databases and health systems are disappearing due to a common CDM (Voss et al. 2015: 557ff.). The use of CDM in other application areas was also investigated. It was found that CDM is well suited for complex analytical queries across systems in the area of active drug surveillance. The performance of these complex queries was proved to be suitable for use in this area (Overhage et al. 2012: 59). Another study investigated whether certain diseases are associated with the month of birth. It was possible to identify 55 diseases of which the incidence was significantly related to the month of birth. 16 of these diseases were named for the first time in context of scientific literature (Bohland et al. 2015: 1044ff.). The OMOP CDM has also been used in the area of comparative effectiveness research. Hripcsak et al. investigated whether the treatment pathways for type 2 diabetes, hypertension and depression differ across countries or are approximately the same. It was found that the treatment pathways differ within and among countries which have participating in the study (Hripcsak et al. 2016: 7333).

3 Methods

In order to fulfil the requirements of the standard concepts in the OMOP model of the OHDSI, new ontologies were integrated into the technical infrastructure of KKR. These are briefly presented in this chapter. In addition, the source data of the operational database of the *Giessen Tumour Documentation System (*ger. Gießener Tumordokumentationssystem = GTDS) will be presented. Furthermore, the manual ETL process is explained in detail. For this purpose, the development environment, lookup table, which are necessary for the transformation step, and the automation of the ETL process are introduced. Finally, the application scenarios that examine the extent to which OHDSI's CDM can be integrated into a cancer registry beyond its initial use, conducting observational research, are described.

3.1 Source Data – Giessener tumour documentation system

The data source used in this paper are all those cancer data that have been documented in a structured form, since establishment of KKR in 2010, and have saved into the KKR registry database. Clinical documentation of individual patients records in KKR is done via the GTDS. It offers a higher structured data input by using templates regarding patient's master data, diagnosis, therapies, tumour boards, treatment course within the framework of a clinical tumour registry (Medizinische Uni Giessen 2020). This involves the representation of individual data in an abstracted form with the aim of data comparability and the provision for quality management and guarantee quality standards in clinical care. For instance, it is easy to investigate whether a tumour board was held prior to determining the treatment strategy. These findings can be actively incorporated into the quality management of clinical care. Moreover, beyond supporting quality management, it is also possible to perform descriptive and analytical evaluations as part of the basic documentation. Patient collectives can be formed and analysed regarding certain parameters such as age, gender and staging, depending on their centre affiliation in the hospital (Dudeck et al. 1999). The GTDS can also determine organ-specific KPIs as part of the cancer centre certification process and in order to guarantee quality of treatment for the patient. These indicators are defined by the DKG. The audits for the certification of organ centres are carried out by *Onkozert*, an independent service provider (see chapter 3.2.2). Furthermore, it is possible to make individual enhancements to the documentation system. Quantitative and qualitative observations can be added by a customized data entry schema. Supported DKG-centres and modules by GTDS can be seen in Figure 5.



Supported DKG modules

Centre	• Oncology • Visceral oncology Breast • Intestine • Gynaecological	• Skin Cancer • Lung Cancer • Prostate Cancer • Sarcoma
Module	• Paediatric Oncology • Head and neck tumours • Liver • Stomach	• neurooncological tumors • pancreatic cancer • oesophagus

Modes of operation

Intrasys-tem	• report modules • Therapy management • Assessment & Analysis	• Study management • Follow-Up & aftercare
Outside the system	• **Data Import** • HL7-Interface • **Data Export** • ADT/GEKID • Melanoma registry	• GTDS-registries • Quality assurance system

Figure 5: supported DKG organ centres and operation principles of GTDS

Workflow tumour documentation

Within the UKE Health information system (HIS) only those patients are listed in a specialized “worklist” that meet certain selection criteria that are mainly 1. diagnosis with "C- or D-codes" based on ICD-10 coding by the central HIS and 2. affiliation to cancer-related departments. Cases imported into the “worklist” and are reviewed by clinical coder for KKR primary responsibilities: the transmission of reporting data, quality assurance and the collection of research data. The subsequent documentation process can be seen in Figure 6. It should be noted, however, that the course of cancer can be heterogeneous, so that the arrangement and frequency of certain events varies from case to case.

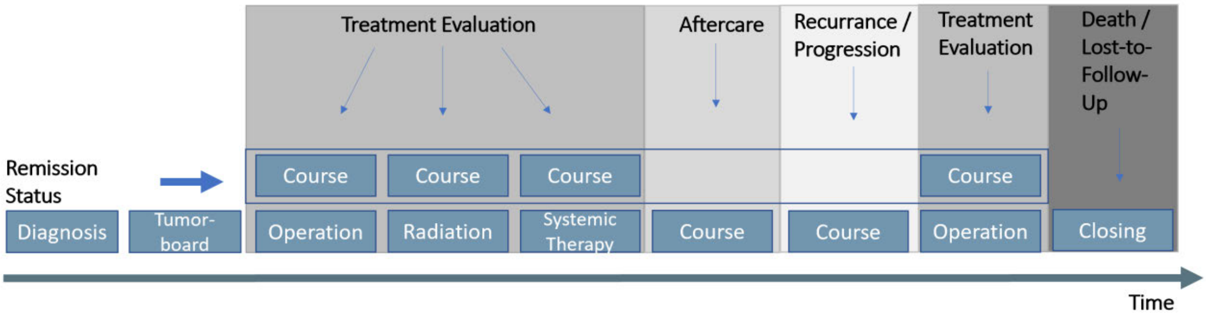


Figure 6: Abstracted possible documentation history of a cancer patient in the GTDS system (following: Medizinische Uni Gießen 2008)

GTDS is a client-server application with an ORACLE database management system as backend. A frontend is provided by an ORACLE-forms and a web-application. This work focuses on the ORACLE database system and the GTDS database embedded in it, since this serves as the main data source for the ETL process carried out in this work. The relational GTDS database has 422 tables that are related by primary and foreign keys. For the correct querying of data, a deep understanding of cardinality of the tables is essential. Primary and foreign keys must be connected correctly to avoid 1. an endless loop and 2. duplicate data entries.

In the present work an attempt was made to transfer individual cancer data, which is collected via GTDS, into the OMOP CDM v.6.0 in order to, on the one hand, extend the existing ontologies in the source system and actively use them as parameters in the formation of patient collectives and, on the other hand, to integrate the collected cancer data into research projects of the OHDSI collective with less costs. Uniform standardized data representation, e.g. by XML schemas, exist for Germany (see ADT/GEKID chapter 3.3.1), but their inclusion in translational research involving other countries is only possible to a limited extent due to the different representation and integration of cancer related ontologies. The transfer of cancer data into a unified format and the exchange of these for research purposes in Germany, has been strongly promoted in recent years. As a next step, the cancer data of the UCCH will be transferred into a worldwide standardized data representation network (OMOP CDM v.6.0) to be used in global research projects.

3.2 Vocabulary

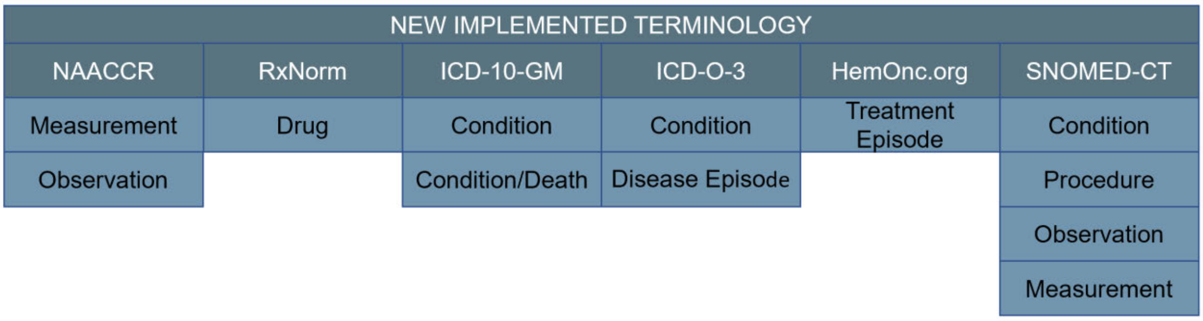
To make the best use of knowledge and to make this knowledge usable in Artificial Intelligence systems, knowledge should be embedded in a common vocabulary with uniform definition of the knowledge transfer by defined entities, classes, relations or attributes which relate to each other to avoid semantic heterogeneity. An ontology provides the relational framework for a common vocabulary and makes shared knowledge possible. According to Gruber "an Ontology is an explicit specification of conceptualization" (Gruber 1993: 199). The representative vocabulary is bound in objects, which are defined by certain relation types or axioms (ebd.: 199).

The OMOP CDM v.6.0 has a vocabulary-driven ontology-based design. Ontologies provide the systematic homogenous representation of heterogeneous data, in medical context (Haendel et al. 2018: 1452). Especially with regard to the increasing technologization and digitization in health care and medicine, an adequate and uniform representation of data in order to use them for research purposes is becoming more important (Smith/Klagges 2008: 21). In the OMOP model, ontologies can be included as source concepts, standard concepts or classification concepts. The developers indicate which ontologies are permitted for the respective concept class (see Tab. 1). In this paper, the vocabulary CMS Place of Service, UB04 Type Bill will not be further explored, because this work aims to map terminology primarily used in cancer setting. If SNOMED-CT is allowed as standard for the target domain to be implemented, it is also integrated into the data model as a standard. For the oncology module the vocabulary ICD-O-3, HemOnc.org and NAACCR was used. In the Domain Drug, RxNorm is designated as standard and ATC as source concept.

Domain	for Standard Concepts	for source concepts	for classification concepts
Condition	SNOMED-CT, ICD-O-3	SNOMED Veterinary, ICD10	MedDRA
Procedure	SNOMED-CT, CPT4, HCPCS, ICD10PCS, ICD9Proc, OPCS4	SNOMED Veterinary, HemOnc.org, NAACCR	None at this point
Measurement	SNOMED-CT, LOINC	SNOMED Veterinary, NAACCR, CPT4, HCPCS, OPCS4, PPI	None at this point
Drug	RxNorm, RxNorm Extension, CVX	HCPCS, CPT4, HemOnc.org, NAAACCR	ATC
Device	SNOMED-CT	Others, currently not normalized	None at this point
Observation	SNOMED-CT	Others	None at this point
Visit	CMS Place of Service, ABMT, NUCC	SNOMED-CT, HCPCS, CPT4, UB04	None at this point

Table 1: Source and standard concepts divided by Domain (OHDSI, 2020: 63)

The ontologies, which are integrated in the CDM, also vary regarding their application scenarios. For instance, ontologies that are able to classify can be easier used in supervised machine learning algorithms, whereas ontologies that do not have classification are more difficult to incorporate into machine learning algorithms (Kulmanov et al., 2020: 14). However, a basic knowledge towards ontologies in the medical context is not sufficient by itself; the data model also has entities that are related to each other in defined relationships. These relationship types are essential for the implementation, especially the correct integration of ontologies of the inner layer into the outer layer of the CDM, and for the correct application and use of the data model. For this work, the ontologies enumerated in Fig 7 were integrated for the first time (except ICD-10-GM) in the KKR and supplemented by the existing terminologies (ICD-10-GM, ATC) in the source system.



NEW IMPLEMENTED TERMINOLOGY

NAACCR	RxNorm	ICD-10-GM	ICD-O-3	HemOnc.org	SNOMED-CT
Measurement	Drug	Condition	Condition	Treatment Episode	Condition
Observation		Condition/Death	Disease Episode		Procedure
					Observation
					Measurement

Figure 7. New implemented standard/source ontologies by Domain

Figure 7 shows which ontology has been implemented in the respective OMOP domain. In the following, the individual ontologies (except ICD-10-GM, since this is already available as terminology in the source system) and their characteristics are briefly presented.

The *North Association of Central Cancer Registries* (NAACCR) defines cancer registry standards for the structured acquisition of data in North America. NAACCR incorporates existing ontologies and classifications, such as the ICD-O-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 is assigned to specific items, which either act superordinate or are assigned in special schemes, according to the respective cancer entity. Each item has a number of expressions (NAACCR value), which are defined by NAACCR or other organization, such as WHO for ICD-O Ontology (NAACCR 2020).

The *National Library of Medicine* (NLM), which is part of the *National Institutes of Health* (NIH) of the USA, provides information and research services with the aim of making biomedical data in the context of healthcare usable for research purposes and simultaneously gaining access to evidence-based results in the field of biomedical research (NIH 2020). NLM developed and administered the ontologies RxNorm and SNOMED-CT. RxNorm provides a clinical drug dictionary for all those drugs that are approved for the pharmaceutical market of the USA. It ensures unique identifiers for pharmaceutical ingredients or brands. Each identifier has certain attributes, which are defined with a respective relationship type. In addition, the identifiers are arranged taxonomically (NIH 2020).

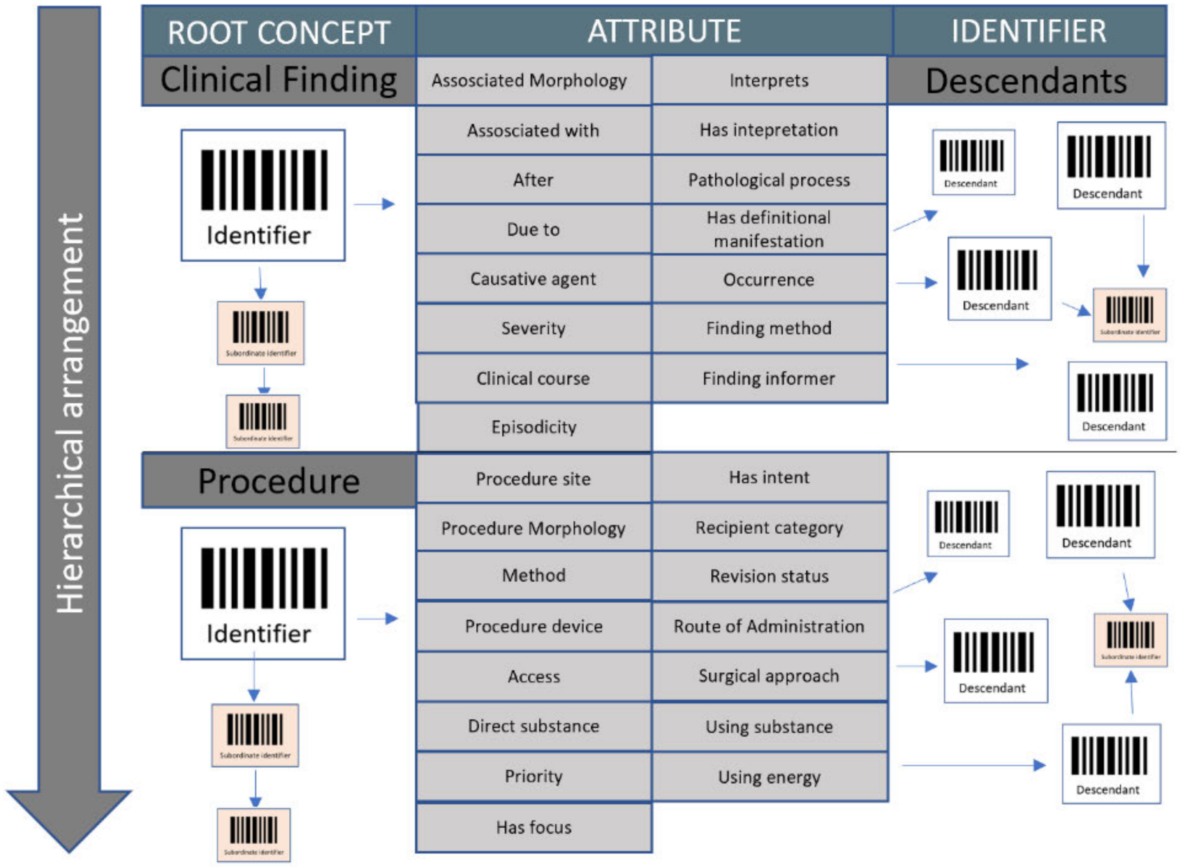


Figure 8: Schematic representation of the SNOMED-CT ontology for Clinical Finding and Procedure concepts

Figure 8 shows the relationship network of SNOMED-CT ontology for the concepts Clinical Finding and Procedure, which are mapped in this elaboration with the source data to the CDM with respect to its standardization. SNOMED-CT ontology has 9 hierarchically arranged concepts, Clinical Finding and Procedure occupy hierarchy 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 ontology. However, it is also possible to infer from the descendants to the root or parent concept. For instance, the concept of triple negative breast tumour (code: 706970001) is associated with the root/parent concept of hormone receptor negative neoplasm (code: 438628005) and the concept of human epidermal growth factor 2 negative carcinoma of breast (code: 431396003). Related concepts can be queried via the attributes “Finding Site” and “Associated morphology” (see Fig. 9). (SNOMED-CT 2020).

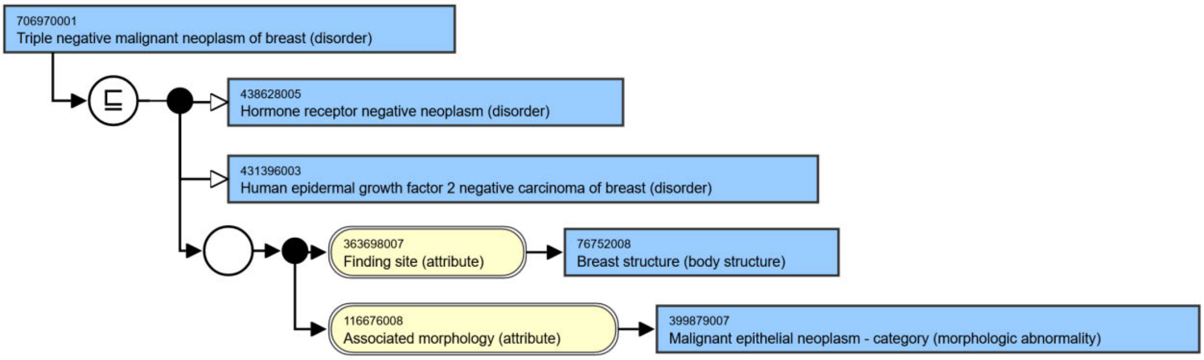
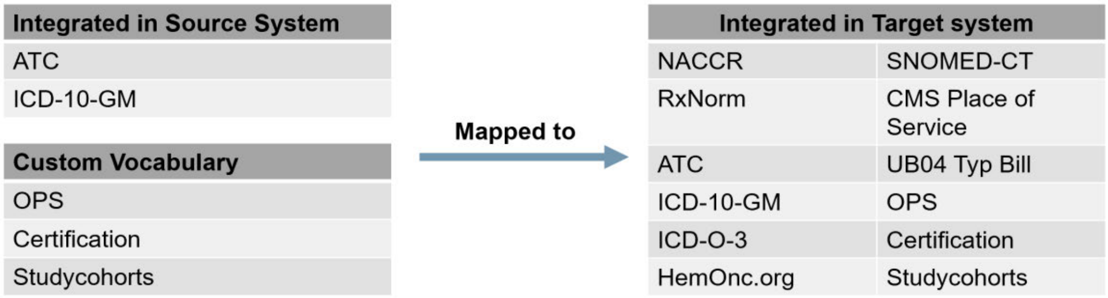

706970001
Triple negative malignant neoplasm of breast (disorder)
⊑
438628005
Hormone receptor negative neoplasm (disorder)
431396003
Human epidermal growth factor 2 negative carcinoma of breast (disorder)
363698007
Finding site (attribute)
76752008
Breast structure (body structure)
116676008
Associated morphology (attribute)
399879007
Malignant epithelial neoplasm - category (morphologic abnormality)

Figure 9: Triple-negative breast tumour SNOMED-CT concept diagram (SNOMED 2021)

ICD-O-3 is a combination of topography and morphology of a tumour. The topography is derived in part from the ICD-10 and has a 4-digit character that covers the range C00.0 to C80.9, which, like ICD-10, specifies tumour categories. (Fritz et al. 2000: 45). The Morphology code of ICD-O-3 specifies the type of cell of the neoplasm and the behaviour. It was derived from *Manual of tumour nomenclature and coding (MONTAC)* and developed and updated over the years by the WHO (ebd.: 3). Today, it has at least a 5-digit character. However, it can be extended by one digit if the degree of differentiation grade or the phenotype of the tumour is further specified (ebd.: 9). The ICD-O-3 is implemented in the CDM in the domains Condition and Episode and links the *condition_occurrence* relation with the disease episodes of the oncology module.

The recording of chemotherapy protocols in the context of cancer treatment is available in the source system in an intern-structured form, which in this context means that the data is only structured and standardized within the source system, but it is not oriented to any homogeneous semantic outside the source system. For the translation of these intern-structured data with respect to the standardized HemOnc.org ontology, the *OncoRegimenFinder* repository of the oncology group is applied. For this purpose, the *OncoRegimenFinder* scripts, which are only available for the 5 version of the CDM, were manually adapted to the 6 model (see chapter 3.2.2). HemOnc.org is a medical Wiktionary. It provides information on treatment regimens, subdivided by disease subtypes and additionally offers information on drugs, interventions but also general information on the treatment of neoplasms (Warner et al. 2015: 337). The HemOnc.org wiki was integrated into CDM to provide a linkage between abstraction of treatment episodes of the oncology module and low-level events of the CDM (Warner et al. 2019: 3).



Integrated in Source System
ATC
ICD-10-GM

Custom Vocabulary
OPS
Certification
Studycohorts

Integrated in Target system	
NACCR	SNOMED-CT
RxNorm	CMS Place of Service
ATC	UB04 Typ Bill
ICD-10-GM	OPS
ICD-O-3	Certification
HemOnc.org	Studycohorts

Figure 10: Schematic representation of vocabulary mapping

Figure 10 shows the schematic process of vocabulary mapping. The ATC and ICD-10-GM terminologies serve as a starting point. In addition, custom vocabulary is integrated into the model to realize the two use cases. Custom vocabulary is only mapped to the SNOMED-CT ontology within the CDM.

3.3 ETL-Process

Operational databases that capture EHR and EMR data are used to support medical staff in clinical practice. Furthermore, these data is used to support healthcare for clinical decision making rather than to using it in translational research (Dennay et al., 2016, p.271f.). In a data warehouse, information from different operational databases flows together. As a result, a modern data warehouse includes a “subject-oriented, integrated, time-invariant, non-updatable collection of data used to support management decision-making processes” and comprehensive research (March/Hevner, 2007: 1031). ETL processes are needed to represent disparate operational databases homogeneously in a data warehouse. It is the basis for the transfer of different data into a unified representation model. The correct implementation of this process is a prerequisite for the success of future data warehouse projects (see Fig. 11) (Vassiliadis et al. 2005: 305f.).

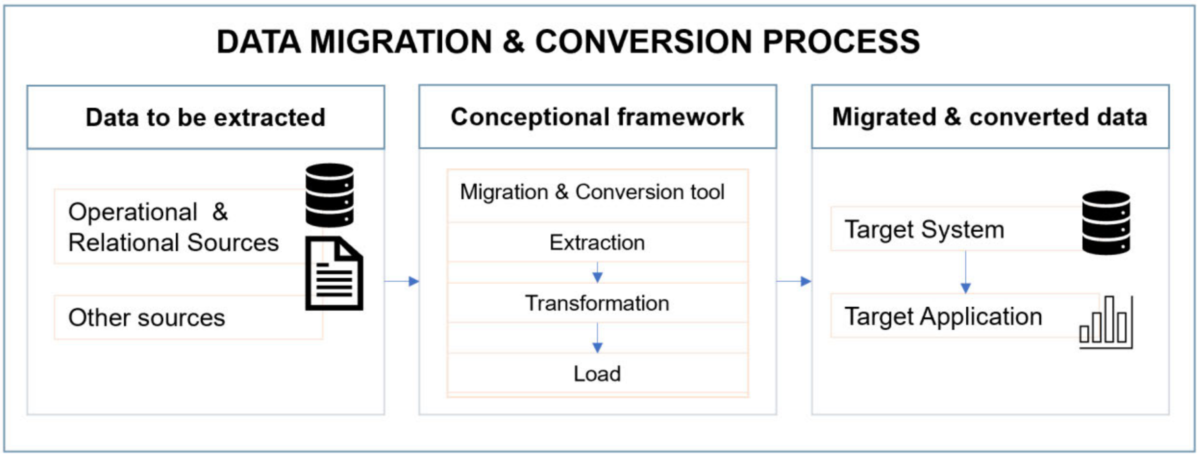

DATA MIGRATION & CONVERSION PROCESS
Data to be extracted
Operational & Relational Sources
Other sources
Conceptional framework
Migration & Conversion tool
Extraction
Transformation
Load
Migrated & converted data
Target System
Target Application

Figure 11: Schematic data migration and conversion process in a medical data warehouse

This chapter describes the ETL process for the extraction of cancer registry data and mapping/transformation into an CDM and the integration into the data warehouse of the UCCH. For the successful realization, the tables of the source database system, which is embedded in a proprietary software (ORACLE), was cloned to an open-source application (PostgreSQL). Since this is only a virtual move of the data and complex transformation steps are not necessary in this ETL process, it was implemented with SQL Server Integration Services (SSIS) toolkit. Whether the realization of an ETL process is implemented tool-based or manually depends on the application scenario (Kimball/Caserta 2004: 10). In this work, the initial ETL process was implemented tool-based from a proprietary system in an open-source application, while the data modelling from the open-source application to OMOP CDM v.6.0 was implemented manually using programming languages SQL, R, PL/pgSQL (see Fig.12).

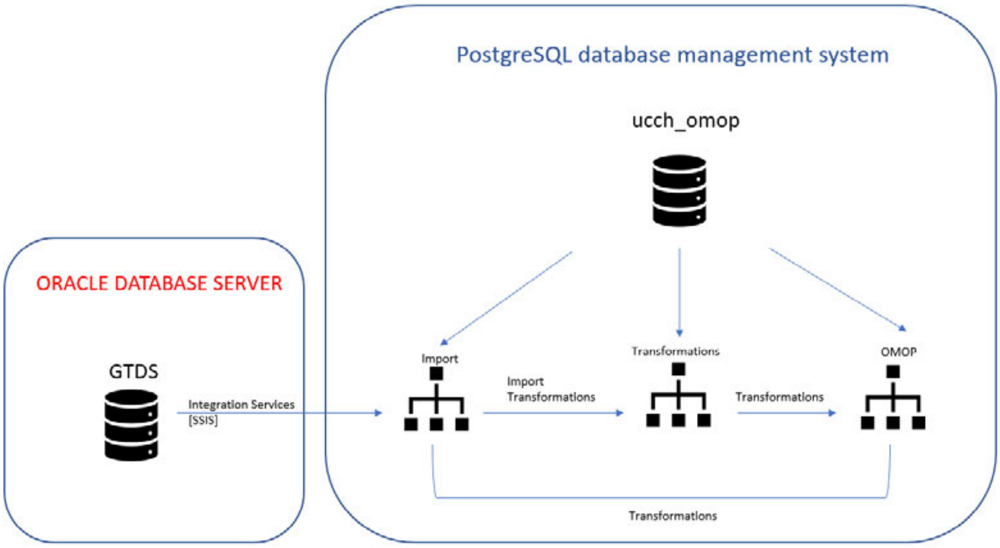


Figure 12: Schematic representation of tool-based and manual ETL-process

3.3.1 Extraction - Data migration

The data that is translated to the CDM originates from the GTDS documentation system (see chapter 3.1). Tables, which are necessary for data integration/transformation with respect to OMOP will be copied to an open-source data management system. This happens for two reasons: 1. the import data is collected centrally in an import schema to ensure integration into the OMOP model from different operational source systems and 2. because the correct data types of the target data warehouse are present in the import system. The selected database management system is *Postgres 12*. Tables, which are migrated to Postgres can be seen in Figure

13 (stored in Database Schema: *import*). In total 34 tables are extracted from the source system and migrated to the open-source application. Based on these 34 tables, three additional transformation scripts (see chapter 3.1.2) are involved in the ETL process to ensure proper data mapping.

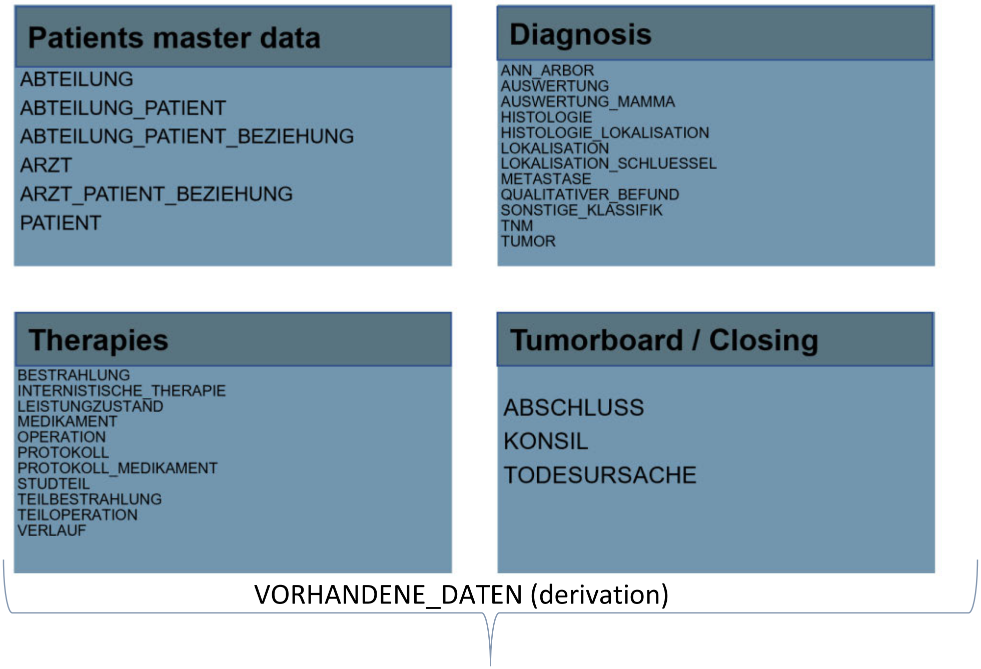


Figure 13: GTDS tables included in the manual ETL-process

3.3.2 Transformations

The main part of data modelling includes the transformation of the GTDS tables into a standardized vocabulary driven OMOP model system. In a first transformation step, GTDS tables were modelled so that they could be integrated into the OMOP ETL process (all_visits, ICDO3_PATIENTS). Furthermore, Lookup Tables (LUT) were integrated into the transformation scripts to guarantee standardization (see Fig. 14).

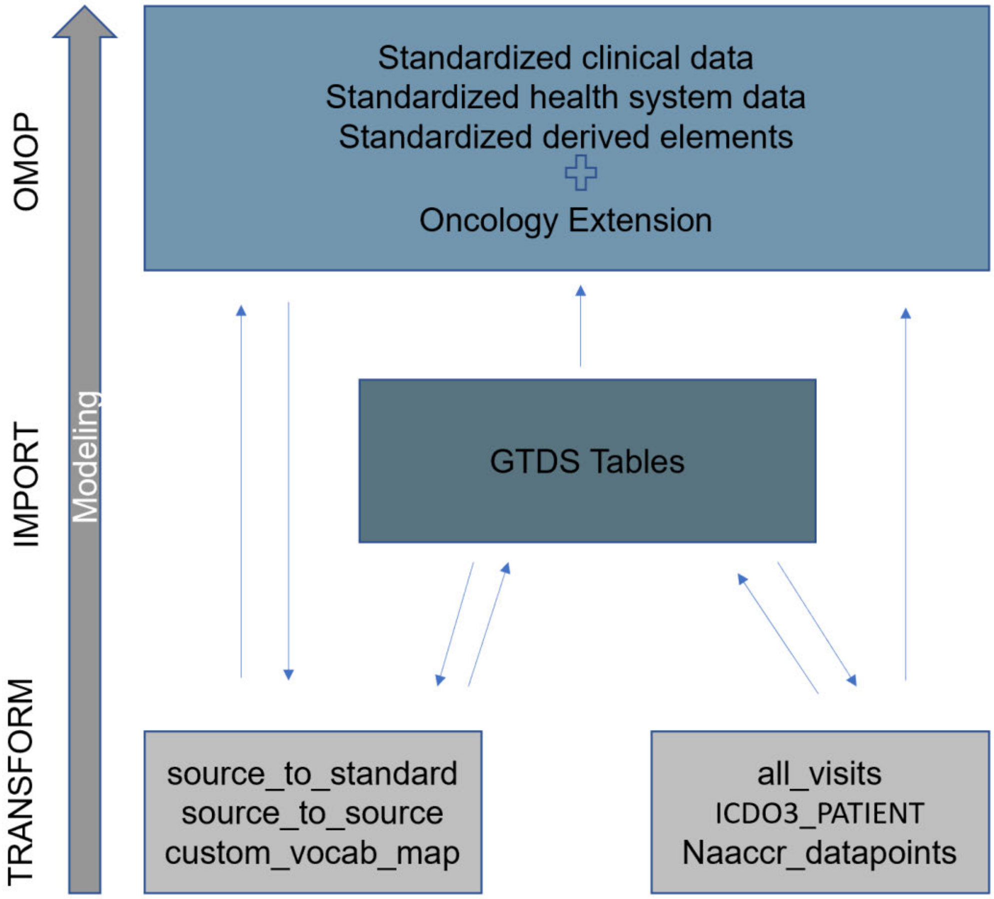


Figure 14: Schematic representation of data modelling

Lookup Tables

Source-to-source-vocab map

For the implementation of standardized source concepts in OMOP tables source-to-source-vocab (Appendix 1) table was used and integrated into transformation scripts.

Source-to-standard-vocab-map

For the implementation of standardized standard concepts in OMOP tables source-to-standard-vocab (Appendix 2) table was used and integrated into transformation scripts.

custom-vocab-map

The implementation of custom standardized vocabulary, which is not included in the OMOP model and therefore must be mapped manually (e.g. OPS) and integrated manually into the relational database architecture of the CDM (table: *concept*, concept_ancestor, concept_relationship, concept_class, vocabulary, source_to_concept_map)

Usagi

Usagi is used for mapping data from source coding systems to the appropriate OMOP standard vocabularies. For the implementation of source vocabulary, which is not present in the CDM, Usagi was developed. The text-algorithm of Usagi provides mapping suggestions regarding defined standard vocabulary in OMOP CDM v.6.0. This tool was used to map the OPS to the standard vocabulary SNOMED-CT, which is declared as standard in CDM for the Procedure domain. Therefore, OPS-Codes were translated via google translate service using a python script and the googletrans package (see Appendix 3) and then loaded into Usagi to fulfil the requirements of text-matching.

Oncology module

For the proper identification of cancer study population there is often the need for additional or other information as in a typical observational study, where the population is defined by exposures and outcomes which are dependent from procedures, diagnostics, or drug exposures. To close this gap in CDM, the oncology module can be integrated into it (OHDSI 2018: 26f.). In order to successfully integrate the oncology module, the CDM is extended by the *episode* and *episode_event* tables (*concept_numeric* is also part of the oncology module but was not implemented within the scope of this elaboration). In addition, the variables modifier_of_event_id and modifier_of_field_id were integrated in the *measurement* table. The *episode_event* table links the disease and treatment abstraction to the underlying clinical events of the CDM. Target vocabulary for oncology module was NAACCR for disease and treatment modifiers, HemOnc.org for the presentation of treatment regimens and ICD-

O-3 for cancer diagnosis. The documentation of chemotherapy protocols in source systems are stored in an intern-structured form, therefore mapping was done via *OncoRegimenFinder* Repository developed by Oncology Working Group of OHDSI collaborative (OHDSI, 2020). This repository identifies patients, which are exposed (dependent on the date difference between drug exposures) with an Antineoplastic Agent (ATC code) and collapses these into appropriate treatment regimens of the HemOnc.org vocabulary (e.g. R-CHOP) (Appendix 6).

To link the CDM with the oncology module, ICD-O-3, based on SEER (= Surveillance, Epidemiology, and End Results Program, e.g. ICD-O-3 SEER Standard = 8520/3-C50.2) coding in USA cancer registries, is needed but not available in this form in the source system. Since histology and topography are stored as stand-alone variables in the source system, Regular Expression was used to ensure ICD-O-3 coding according to SEER (Appendix 7). Disease Episodes are also not present in the source system. Instead, it is possible to assess the disease situation at date of event. A query was developed (Appendix 5), under use of temporary tables, which summarizes these measuring points into time intervals (Complete Remission, Disease Recurrence, Disease Progression). Table 2 shows the first transformation step while Table 3 displays the final transformation step as it is implemented in the target system.

FK_ TUMORFK PATIENT	FK_TUMOR TUMOR_ID	UNTERS _DATUM	LFDNR	GESAMT-BEUR-TEILUNG	PRIMAER-TUMOR	LYMPH-KNOTEN	META-STASE
0	1	12.10.2015 00:00	4	P	K	K	R
0	1	16.11.2015 00:00	5	O	K	K	K
0	1	15.05.2018 00:00	6	O	K	K	K
0	1	15.06.2018 00:00	7	V	K	K	K
0	1	25.06.2019 00:00	8	P	K	K	B
0	1	05.03.2019 00:00	9	O	K	K	K
0	1	24.07.2018 00:00	10	O	K	K	K
0	1	28.06.2017 00:00	11	V	K	K	K
0	1	09.08.2017 00:00	12	P	K	K	R
0	1	24.01.2018 00:00	13	X	K	K	M
0	1	05.12.2018 00:00	14	X	K	K	M
0	1	24.06.2015 00:00	15	X	K	K	M
0	1	08.07.2014 00:00	16	X	K	K	M
0	1	28.08.2014 00:00	17	P	K	K	P
0	1	12.10.2015 00:00	18	P	K	K	R
0	1	13.07.2016 00:00	19	X	K	K	M
0	1	28.06.2017 00:00	20	X	K	K	M
0	1	15.01.2017 00:00	21	P	K	K	P
0	1	08.07.2014 00:00	22	P	K	K	P

Table 2: Assessment of disease course with date events as it is presented in source system

Episode_id	Person_id	Episode_concept_id	episode_start_date	episode_end_date	episode_parent_id	episode_number	episode_object_concept_id	episode_type_concept_id	epsiode_source_value	episode_source_concept_id
4167	00	32528	06.04.2010 00:00	01:38,8	0	1	4237178	32546	0	44505315
78333	00	32529	08.07.2014 00:00	08.07.2014 00:00	0	2	4237178	32546	0	44505315
51179	00	32530	08.07.2014 00:00	28.06.2017 00:00	0	3	4237178	32546	0	44505315
78332	00	32529	28.06.2017 00:00	15.06.2018 00:00	0	4	4237178	32546	0	44505315
92721	00	32677	15.06.2018 00:00	47:39,1	0	5	4237178	32546	0	44505315

Table 3 : Assessment of disease course as it is presented in OMOP CDM v.6.0

Furthermore, NAACCR is not used in European context. Treatment and diagnostic modifiers in cancer context are not stored in a European valid nomenclature. Instead, they are integrated into an internationally valid terminology. This terminology was parsed and transformed to NAACCR nomenclature. The resulting LUT (naaccr_datapoints) was used for the implementation of diagnostic and treatment modifiers in *measurement* and *observation* table of the CDM.

3.3.3 Load

Data was loaded into the target systems on a daily basis using Windows Scheduler, which manages and runs jobs. For this purpose, the different scripts were included in batch files and a trigger was created so that the ETL process is updated every day to ensure an up-to-date data basis (see Tab. 4). The first step starts, in transform schema, the necessary data transformation process to fulfil the target system requirement for implementing the CDM (Transformation_etl.bat, transformation schema). Afterwards the scripts are read in, which are necessary for the implementation of the CDM (Opmopetl.bat, omop schema). To meet the system requirements for the oncology module, the Oncology WG's *OncoRegimenFinder* repository was used to detect treatment regimen, based on ATC, in the source system. The R-scripts were also stored in a batch file and included as a task in the Windows task scheduler (onco_regimen.bat, transform schema). Next, it runs the queries to create the NAACCR measurement points (Naaccr_etl.bat, transform schema), which are included as modifiers in the oncology module in the *measurement* and *observation* tables. After the system requirements for the oncology module have been met, the process for implementing the oncology module (Onco_module_etl.bat, omop schema) is started.

Rank	Script	Description
1	Transformation_etl.bat	Data transformation to fulfil target system requirements
2	Omopetl.bat	Implementation of OMOP CDM v.6.0
3	Onco_regimen.bat	OncoRegimenFinder R-Script for the proper identification of treatment regimens
4	Naaccr_etl.bat	Derivation of NAACCR items from GTDS
5	Onco_module.bat	Implementation of oncology module

Table 4: Load Process

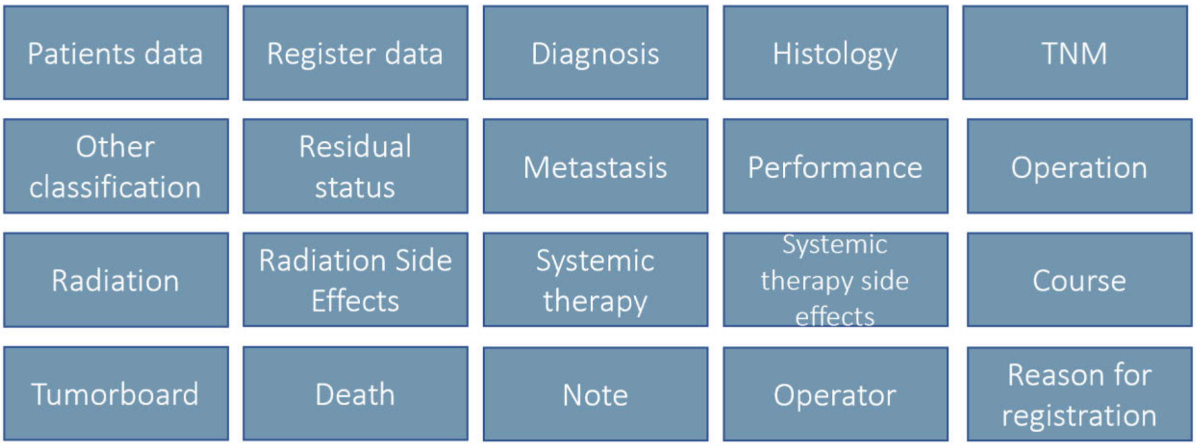
3.4 Evaluation of OMOP CDM v.6.0

After completion of the ETL process, data modelling was evaluated using various application scenarios. These specific application scenarios go beyond the initial development of the model. The aim was to cover the working areas of the clinical cancer registry and to determine in which application scenario the model is best suited. The following chapter describes two different use cases which try to present the real application scenarios of the KKR: reporting and quality assurance (see chapter 2.1), which are described afterwards.

3.4.1 Transmission of ADT/GEKID on OMOP CDM v.6.0

Since 2014, the HmbKrebsRG has been actively implemented in the federal state of Hamburg. Physicians, medical care centres and hospitals are obliged to transmit cancer information to the respective state registry within a period of 8 weeks. The federal state registries in turn transmit their data to the *Robert-Koch-Institute*, which includes it in its health reporting. The ADT/GEKID record provides the framework for the transmission of cancer data. In addition to the basic data set, there are also specific modules that query additional items, depending on the respective diagnosis (ADT, 2020). The transmission of the reporting data to the respective national register is done via Extensible Markup Language (XML) in a specified XML-Schema (XSD, see https://basisdatensatz.de/download/ADT_GEKID_v2.2.1.xsd). Transmitting cancer data in a uniform defined XSD allows data from heterogeneous source systems to be represented in a consistent manner. Both approaches, XSD and CDM, provide a unified framework for knowledge representation and data integrity from different source systems (Klein et al. 2001: 6).

In this paper, only the mapping of data according to the ADT/GEKID basic schema on the OMOP model is examined. It was examined whether the translation of the 20 (see Fig. 15) possible ATD/GEKID categories with their associated items can also be determined by means of the OMOP data model in addition to XSD. This is primarily an application scenario for data exchange within the CCC in the context of research projects. The transfer of the ADT/GEKID dataset into the OMOP vocabulary could greatly facilitate data exchange and cross-institutional analysis of cancer data in Germany, as this form of cancer data representation is available at all CCCs. In this paper the mapping of ADT/GEKID to OMOP CDM will be investigated in a theoretical framework.



Patients data	Register data	Diagnosis	Histology	TNM
Other classification	Residual status	Metastasis	Performance	Operation
Radiation	Radiation Side Effects	Systemic therapy	Systemic therapy side effects	Course
Tumorboard	Death	Note	Operator	Reason for registration

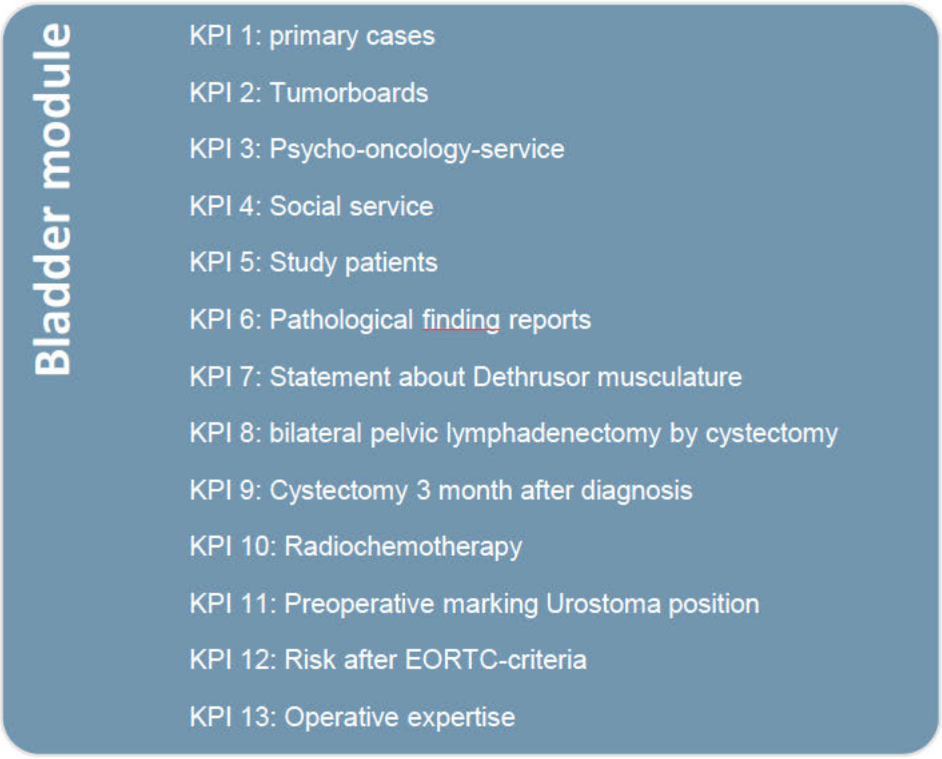
Figure 15: ADT/GEKID Categories

For this purpose, each category of ADT/GEKID was examined individually and a score for the mapping rate was determined. Furthermore, a total mapping rate was determined after the evaluation process. A weighting of the different items was not made. Possible scores were Yes, No and NULL. The value NULL was given if the mapping of this item would be possible but would require an individual extension or an extension provided by OHDSI of the CDM (except oncological module).

3.4.2 Certification

For the examination of the suitability of the CDM for the determination of organ-specific characteristic numbers in the context of the DKG certification, OMOP CDM v.6.0 is used to calculate KPI defined by the DKG for the entity urinary bladder model (see Fig. 16). The urinary bladder module is part of the uro-oncological centre certification, in which, in addition to the prostate, at least one other urological organ (kidney, bladder) is assessed as part of the DKG certification. It is also possible to include the bladder module within the framework of an oncological centre (Onkozert, 2020). Besides the determination of organ-specific indicators, the certification process also includes other areas such as the inspection of wards and the provision of specific treatment measures using technical equipment. However, since the responsibility for the prerequisite of these certification characteristics is not part of the KKR, only the provision of data is discussed here. The data collection form of the bladder module consists of 13 KPIs, which require a numerator/denominator specification. Furthermore, a target specification is defined. If this target is not met, a written explanation of this data deficit is required. The calculation of the KPIs to meet the certification requirements of DKG is determined individually by each institute. The use of the documentation system and the operational databases, which are integrated into the technical infrastructure of the respective institute differ. This means that the data representation within the operational database or the data warehouse varies from institute to institute. As a result, each institute must write individual queries to determine the KPIs specified by the DKG. This use case examines the extent to which the OMOP CDM can be used to provide common certification queries that can be used independently of the documentation system, across institutions that have their centres certified as part of DKG certification and included the OMOP model into their medical data warehouse.

All 13 KPIs were determined using the OMOP model. The KPIs calculated by the OMOP are compared with the KPIs from the source system and checked with regard to the deviation, assuming that the source system reflects the actual values. Custom defined vocabulary had to be integrated into the OMOP model to calculate KPIs of the bladder module. Therefore, vocabulary which includes features of the GTDS certification vocabulary and features for study participation of cancer patients in the GTDS were integrated as custom vocabulary in metadata of OMOP CDM v.6.0.



Bladder module

KPI 1: primary cases
KPI 2: Tumorboards
KPI 3: Psycho-oncology-service
KPI 4: Social service
KPI 5: Study patients
KPI 6: Pathological finding reports
KPI 7: Statement about Dethrusor musculature
KPI 8: bilateral pelvic lymphadenectomy by cystectomy
KPI 9: Cystectomy 3 month after diagnosis
KPI 10: Radiochemotherapy
KPI 11: Preoperative marking Urostoma position
KPI 12: Risk after EORTC-criteria
KPI 13: Operative expertise

Figure 16: KPIs for DKG bladder module

4 Results

Following, the results of the ETL process are presented, with focus on the implementation of the standardized vocabulary into the CDM. The data modelling section highlights the integration of source data with respect to OMOP standardization from different points of view. The model was also tested in terms of applicability. For this purpose, two use cases were defined, which were presented/calculated via the OMOP model. The results of the use cases are shown at the end of the chapter.

4.1 Data modelling

This section highlights the mapping of the vocabulary. For this purpose, the information transfer from the source system to the target system is shown. Furthermore, it is analysed to what extent the terminologies integrated in the source system (OPS, ICD-10-GM, ATC) could be mapped to the standardized ontologies integrated in the OMOP model. Intern-structured data (data entries that are not defined by nationally applicable terminologies but have a valid standardization within the source system), was integrated into the data model as custom vocabulary in order to meet the requirements for the first and second use case. Therefore, custom vocabulary is considered only to a limited extent, especially in the analysis of the vocabulary mapping. In addition, it was investigated to what extent the source and standard concepts in the respective domains could be mapped to the OMOP standardization. Next, concept mapping is presented. It is considered how many distinct concepts of the target ontologies (RxNorm, SNOMED-CT) could be covered with the source data. It is also considered manual mapping via the Usagi software, which was used to map OPS procedures to the SNOMED-CT ontology. Furthermore, the integration of the oncological module is presented. Especially the integration of disease and treatment episodes will be presented because these dimensions are crucial for the analysis of cancer data. In this context also the implementation of the HemOnc.org vocabulary and episodic modelling of cancer will be discussed.

Information transfer

For the evaluation of information transfer, the area of standardized clinical data represented by the domains Condition, Drug, Procedure, Measurement, and Observation within the OMOP Model will be examined in more detail. Figure 17 shows the total number of data entries in the source and target system divided by the domains of the OMOP model. During the transfer of the source data into the standardization of the target system, there is a loss of information in the areas of Condition, Measurement, Observation and Procedure ($\bar{x}$ = 11850). In the drug domain there is a gain on information. This is due to the fact that the ATC ontology, which was mapped into the OMOP model as standardized source concepts, partially has a 1:n mapping with respect to the RxNorm ontology, which is implemented as standardized standard concepts in the drug domain.

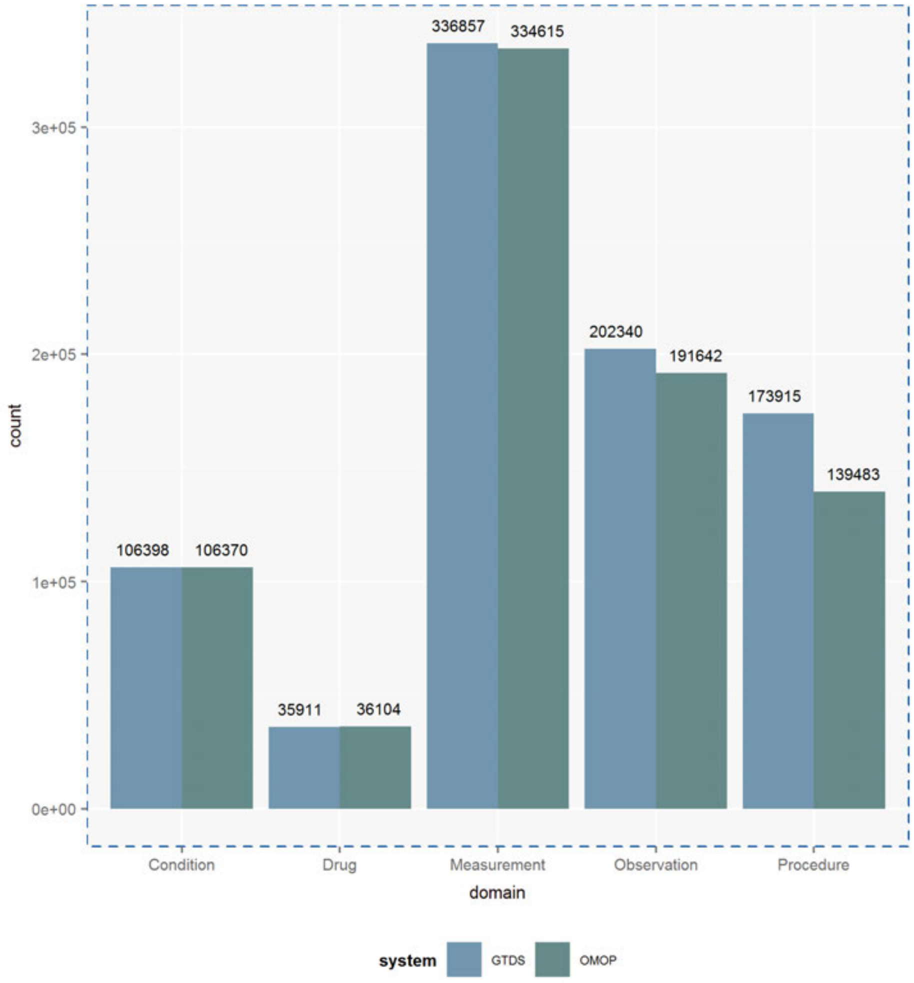


Figure 17: Total amount of data entries subdivided by domain OMOP/GTDS

Translation rate of terminologies in source system to corresponding Ontology in target system

During the data integration process, several terminologies were successfully integrated into the OMOP data model. Existing terminologies in the source system were used as a starting point for the mapping of additional ontologies, which are integrated into the CDM. The terminologies ATC and ICD10-GM were adapted to the standardization of the CDM and intern-structured data (Certification, Studycohorts) from the source system were integrated into the CDM system as custom vocabulary. Since the terminologies, which are integrated in the source system (OPS, ICD10-GM, ATC) are the starting point for data mapping to corresponding ontology in the CDM that are not present in the source system, the translation rate of the source ontologies is of particular importance to evaluate the implementation of new ontologies in the context of their data basis. In the domain of drugs 93% of the source data which are describing drug/drug ingredient were translated successfully to the standardized ATC ontology included in CDM (n= 33421). Starting from this, 33362 events have been mapped to the RxNorm ontology, which corresponds to a transfer rate from source ontology to standard ontology in drug domain of approx. 100%.

VOCABULARY	SOURCE SYSTEM	TARGET SYSTEM	translation_rate (≈%)	not_translated (≈%)
OPS	125691	80349	64%	36%
ATC	35832	33421	93%	7%
ICD-10GM	53175	53171	100%	0%

Table 5: Translation rate of ontologies, which are integrated in source system

Custom defined vocabulary (except OPS) is not displayed in table 5. They were only implemented to create the basic prerequisite for the representation of the use cases (chapter 3.4.1, 3.4.2, 4.2.1, 4.2.2). The ICD-10-GM, which is used in GTDS to code diagnosis, was mapped to the ICD-10-GM ontology, which is integrated in the OMOP vocabulary in a standardized way as source ontology in the Condition domain. With the exception of 4 data entries, all data, which describes the coding of the diagnosis in the source system could be mapped to the ICD-10-GM ontology stored in the CDM vocabulary.

Concept mapping

The source data could be transferred to 1936055 concepts of the OMOP model. This data pool of concepts feeds into most of the OHDSI collaborative applications because there is standardization here, allowing cross-institutional comparison. From the integrated concepts, a total of 2744 distinct source and standard concepts can be derived. The most distinct concepts are in the *Condition* domain, in which, at the time of the work, the deaths codes in are not yet integrated. Therefore, it can be assumed that this value will increase significantly after the integration of deaths. Most of the integrated concepts are in the Measurement domain, since in this domain primarily the NAACCR ontology is used, which is not available in Europe and a data mapping of the source data is more difficult, these concepts are only assigned to 39 distinct concepts. Looking at the episode concepts of the oncology model, the standard concepts denote the respective episode (disease first occurrence, disease remission, disease progression, disease recurrence, treatment episode). In the domain of the disease episode, the ICD-O-3 diagnosis codes are specified as source concepts. A total of 440 distinct ICD-O-3 concepts are available in the disease episode domain. In the treatment episode, the source concepts are presented by the HemOnc.org. The

OncoRegimenFinder repository was able to derive 35 distinct HemOnc.org concepts from the source data.

The domains Person, Location and Care site have no concepts integrated, because these are not available in the source system (for instance race or ethnicity of a person). Therefore, only the absolute number of data entries in these domains is listed.

Domain	Concepts/n	Distinct standard concepts	Distinct source concepts	Distinct Relationships
Condition	162004	619	740	25
Drug	66783	124	122	20
Procedure	125032	582	0	41
Measurement	706179	26	13	8
Observation	356920	31	3	7
Disease Episode	148588	4	440	3
Treatment Episode	9338	1	35	15
Visit Occurrence	298702	4	0	3
Person	49681	-	-	-
Location	12445	-	-	-
Care Site	383	-	-	-
Total	**1936055**	**1391**	**1353**	**122**

Table 6: Total count, distinct count of integrated concepts/data entries

Concepts are assigned to relation types in the OMOP model, which could be queried via the *concept_relationship* table in the standardized vocabularies area. This linking of relationship types makes it possible to query additional information of a concept without this information is being available in the source system. For instance, Concept_id: 21603761 (= Bevacizumab - ATC) is a (Relationship_id) Concept_id: 21603754 (= Monoclonal antibody – ATC). The information that bevacizumab was administered is derived from the source system, whereas the information about the drug class is provided by the OMOP model. Relationship types of the OMOP model thus, lead to the fact that additional information can be derived from the source data. In the context of this work, the source data were linked to 122 distinct relation types.

relationship_id	count	Percent [%]
Maps to	6070	14
Is a	5733	13
Mapped from	4764	11
Has asso morph	3982	9
Has finding site	3891	9
Has Histology ICDO	3537	8
Has Topography ICDO	3537	8
ICDO to Schema	3537	8
ICDO to Proc Schema	3363	8
Subsumes	1061	2
Total	43545	100%

Table 7: Total count top 10 relationship types of integrated concepts

Table 7 shows the absolute number of integrated relationship types. A total of 43545 concepts are associated with a Relationship type. The relationship type “Maps to” occurs most frequently with 14%. “Is a” and “Subsumes” relationship types provide information about a hierarchical arrangement. The parent concept is always queried for these relationship types. However, if the descendants of the concept are to be accessed, this can be done via the *concept-ancestor* table instead of *concept_relationship* table.

Mapped Vocabulary/Concept by Domain

The number of successfully mapped data items divided by their corresponding domain in the OMOP model is shown in Tab. 8.

domain_id	vocabulary_id	n	% by domain
Measurement	NAACCR	334550	100.00%
Observation	NAACCR	165278	86.21%
Procedure	SNOMED	124951	48.12%
Procedure	OPS	80349	30.95%
Meas Value	NAACCR	76054	100.00%
Visit	UB04 Typ bill	73584	100.00%
Condition	SNOMED	59014	36.43%
Condition	ICD-10-GM	53171	32.82%
Condition	ICDO3	49821	30.75%
Drug	ATC	33451	50.07%

Drug	RxNorm	33362	49.93%
Procedure	GTDS_internal_therapy	35704	13.75%
Observation	Certification	25659	13.38%
Procedure	GTDS_Radiation	18637	7.18%
Regimen	HemOnc	14272	100.00%
Observation	Studycohorts	626	0.33%
Observation	SNOMED	160	0.08%

Table 8: Number of translated data entries depending on their implemented ontology and domain

The Measurement domain, which primarily integrates the NAACCR ontology to identify treatment and diagnostic modifiers in oncology module, has the most total amount of standardized events (n=334550). However, not all events could have been transferred successfully to the standardization of the OMOP model. Depending on the domain, some of the source data was implemented several times in the corresponding domain, keeping the standard concepts defined by the developers. Thus, the ontologies ICD-O-3 and ICD-10-GM were included in the domain as standardized source concepts and have been mapped to the corresponding standardized standard concepts of SNOMED-CT ontology. Nevertheless, the standardized standard concepts, represented through SNOMED-CT ontology (n= 59014) are only slightly higher than the total number of standardized source concepts, represented by ICD-10-GM (n= 53171) and ICD-O-3 (n=49821) ontology, because only 11.7% (n=5841) of ICD-O-3 events could be mapped to standardized standard SNOMED-CT concepts. Whereas for ICD-10-GM almost all events (≈100%) could be mapped to the standardization of the SNOMED-CT ontology.

Figure 18 illustrates source and standard concepts which have been integrated into the standardization (~mapped) of the model and those concepts which could have been integrated but were not transferred into the uniformly standardized format (~not mapped) divided by domain. For the analysis, the standardization of standard concepts and the standardization of source concepts were reviewed. An exception are custom concepts, which are always assigned the value 0 as standardized source concepts. Therefore, these data entries are only checked whether these are assigned to a standard concept with the respective standardization. Since most of the data entries of the custom vocabulary, which are integrated into the CDM, assigned to a unique standard concept pedant, this leads to higher match scores in domains with a lot of custom vocabulary. In total ≈94.2% of the translated data could be mapped to the standardization of the OMOP model. The best results were achieved in the domain of

Observation and Measurement (100%). At this point it should be noted, that in the Observation and Measurement domain many custom vocabulary was integrated and only the NAACCR ontology was implemented as standardized source and standard concepts (see chapter 5.1).

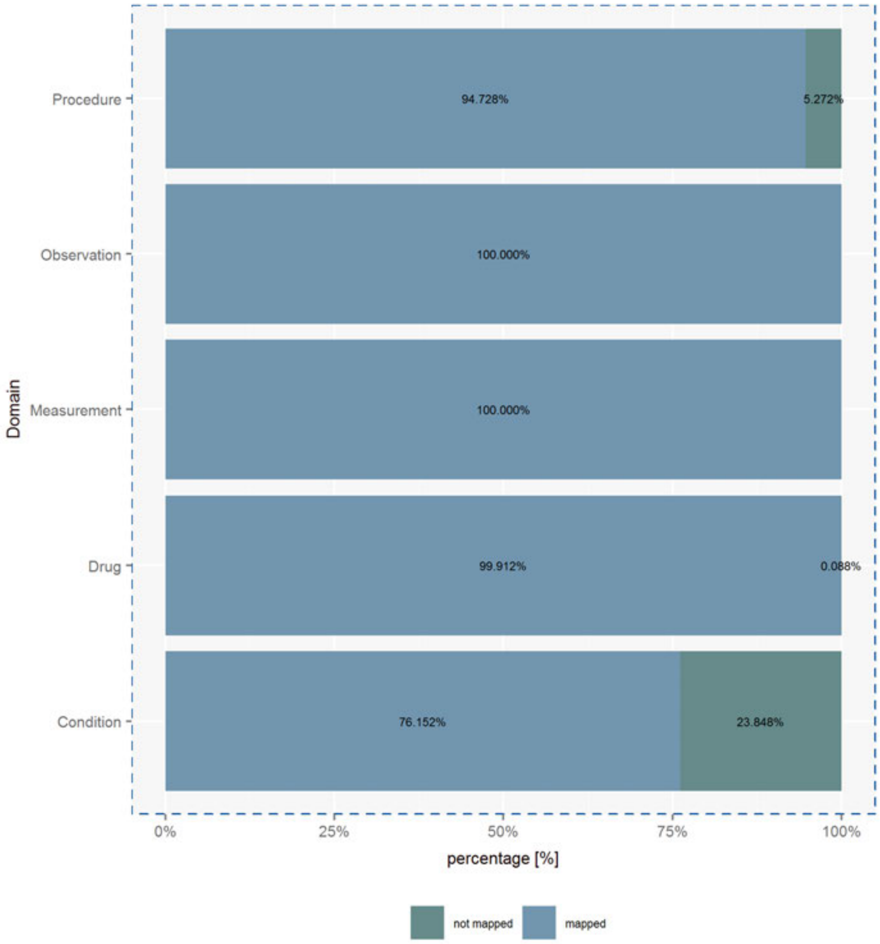


Figure 18: Mapped and unmapped data entries depending on their domains

Usage rate of mapped ontologies in CDM with the data from the source system

SNOMED-CT

The SNOMED-CT ontology has been integrated into the Condition, Observation and Procedures domains. A total of 184125 data entries could be assigned to a standardized SNOMED-CT concept. The exhaustion of distinct SNOMED-CT concepts in the domains Clinical Finding and Procedure, however, amounts to only 0.54%. This reflects an absolute distinct mapping of 1546 (a total of distinct 284493 SNOMED-CT concepts in CDM Vocabulary) codes from the source system to the corresponding SNOMED-CT ontology in the CDM.

RxNorm

In the drug domain, 124 different distinct could be transferred from ATC ontology to the corresponding RxNorm concepts. This corresponds to a usage rate (implemented RxNorm codes in CDM Ontology: n=287928) of 0.039%.

OPS-SNOMED-CT Mapping via Usagi

Usagi

During this work, the OHDSI vocabulary group included the OPS ontology in the CDM vocabulary. The mapping was done exclusively by text-matching to the existing CDM vocabulary for the domains *Device* and *Procedure*. Nevertheless, both OPS mappings, manually mapped and text-based-map, were applied and compared and let to the result, that the matching score for the domain *Procedure* with the manual mapping achieved a better result compared to the OPS mapping of the OHDSI Vocabulary group (Matching Score Procedure Domain manually mapped = 94.7%, Matching Score Procedure Domain by Vocabulary Group = 91.2%). Therefore, it was decided to use the manually mapped OPS codes for implementing OPS Codes from the source system to appropriate standardized standard SNOMED-CT concepts in the procedure domain. In the future, however, this should be changed, as the application possibilities for the ontologies integrated in CDM vocabulary are better compared to custom vocabulary. The OPS catalogue of the KKR contains of 5344 different codes, which are included in the source system in different frequencies. In contrast to other terminologies not the entire OPS catalogue has been mapped to the SNOMED-CT ontology. In this paper only those codes, which are frequently present in the source system has been processed. OPS codes designating a biopsy are more common in the source system than those used to describe, for example, the creation of an aortocoronary bypass due to treatment course. As a result, it was tried to map primarily those OPS codes representing oncological treatments. A total of 944 distinct OPS codes were mapped to the SNOMED-CT ontology. Table 9 lists the most frequently occurring OPS codes in the target system with their corresponding SNOMED-CT counterpart. The top 5 mapped OPS codes cover 8468 events, representing approximately 10.5% of the OPS events mapped to the SNOMED-CT ontology. The OPS codes were mapped with an 1:n relationship to the SNOMED-CT ontology. This means that one OPS code has a corresponding SNOMED-CT pedant, at the same time a SNOMED-CT code can be assigned to several OPS codes. In total, 80349 events with OPS encoding could be translated to the SNOMED-CT ontology in the target system. This corresponds to a mapping rate of OPS codes from the source system to the target system of 64%.

In addition, the non-standardized terminologies used in the source system for radiological and drug therapies were added to the model as custom vocabulary and then also mapped to SNOMED-CT ontology. In the area of drug therapies only 38 data entries were not mapped to standardized standard concept, representing SNOMED-CT in the procedure domain of the CDM. The outcome of mapping radiological and drug treatments has been more successful compared to surgical procedures (not mapped OPS Codes in Procedure Domain: n= 14180).

OPS-Code	**SNOMED-Code**	**n**	**OPS-Description**	**SNOMED-CT-Description**	**percent [%]**	**cum sum [%]**
1-494.31	277590007	2166	(Percutaneous) to biopsy other organs and tissues of control by imaging methods: Mamma: By punch biopsy without clip flag of the biopsy region	Imaging guided biopsy	2.7	2.70
5-010.00	25353009	1862	Craniotomy over the Dome: craniotomy (cap): dome	Craniotomy	2.31	5.01
5-401.11	443497002	1567	Excision of individual lymph nodes and lymph vessels: axillary: With Radionuklidmarkierung (sentinel lymphadenectomy)	Lymphadenectomy of sentinel lymph node	1.95	6.96
5-870.a2	1054971000000105	1498	Partial (breast-conserving) excision of the breast and destruction of breast tissue: Partial resection: defect coverage by mobilization and adaptation by more than 25% of the breast tissue (more than 1 quadrant)	Primary vertical reduction mammoplasty with nipple graft and excision of more than 1500g of tissue	1.87	8.83
5-021.2	120075000	1375	Reconstruction of the meninges: duraplasty, frontobasal	Brain reconstruction	1.71	10.54

Table 9: Top 5 mapped OPS codes in CDM

Oncology Module

Due to the oncological data model extension a total 91452 episodes related to treatment and diagnosis could be integrated into the oncological module. The best data transmission was achieved with the concept of initial diagnosis (Fig. 19) This results in a translation rate of 92.4% (source initial diagnosis= = 53199). The lack of an outcome of 100% mapping regarding initial diagnosis concept is due to the fact that the ICD10-GM terminology is used for coding the cancer diagnosis in source system, whereas ICD-O-3 is used in the target system. The ICD-O-3 is not available as a stand-alone variable in the source system and was formed by Regular Expression using the localization and side annotation.

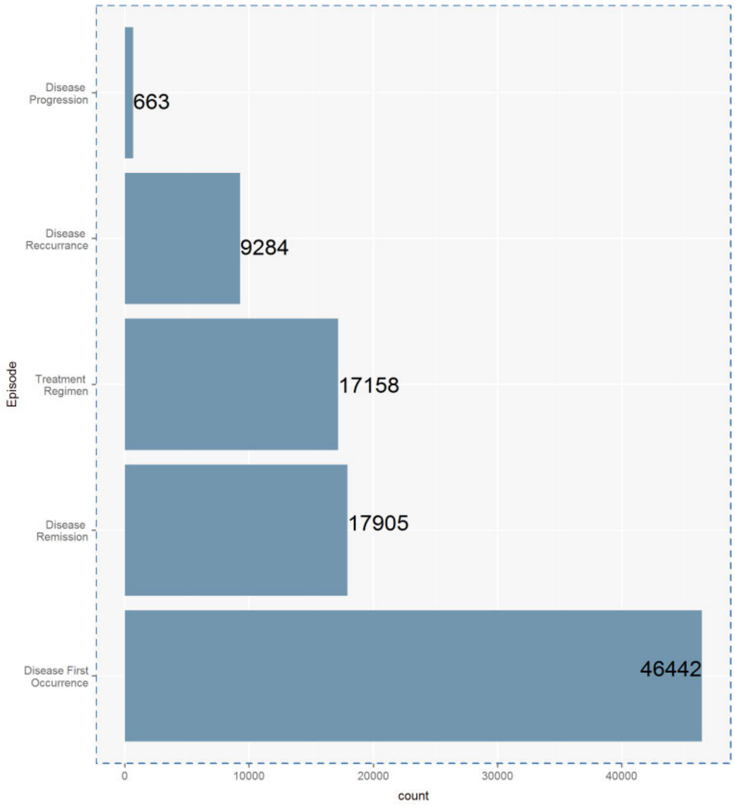


Figure 19: Implemented treatment and disease episodes in Episode Domain

Due to the existing representation of remission status at specific date events instead of time spans in the source, less data was integrated into the target model compared to source system. A total of 137224 date events are used to assess the disease in source system. From these, 27852 time intervals could be derived and assigned to the disease episodes of the oncology module. As soon as the concept of the stable disease is included in the CDM vocabulary, it can be assumed that the number of derived time intervals will increase further. Fig. 19 shows the total amount of integrated episodes in the CDM subdivided in their categories.

The HemOnc.org vocabulary was used to display the treatment episodes. A total of 17158 treatment regimens could be identified through *OncoRegimenFinder* Repository (see chapter 3.2.2) and then be transferred in the target system, however, 2886 data entries could not be mapped into a standardized standard concept, representing HemOnc.org in the Episode domain of CDM Oncology module. At the time of modelling, the source system contained 20146 data entries to specify the applied therapies, those therapies, which are not associated with drug administration (watch & wait, palliative care, etc.) where excluded from this analysis because they have no valid ATC value.

DESIGNATION	**regimen**	**n**	**prop. [%]**
Femara (Letrozol)	Letrozol (GTDS)	820	4.1
Tamoxifen	Tamoxifen (GTDS)	797	4.0
TACE (Doxorubicin + Lipidol)	Doxorubicin, Lipidol (GTDS)	549	2.7
Avastin (Bevacizumab)	Bevacizumab (GTDS)	540	2.7
Mitomycin Frühinstillation	Mitomycin (GTDS)	527	2.6
ddA	Doxorubicin (OMOP)	1062	7.4
Carboplatin and Paclitaxel	carboplatin,paclitaxel (OMOP)	855	6.0
Cisplatin monotherapy	Cisplatin (OMOP)	701	4.9
Tamoxifen and OFS	Tamoxifen (OMOP)	678	4.8
Letrozole monotherapy	Letrozole (OMOP)	627	4.4

Table 10: Top 5 treatment regimen in source system (GTDS) and target system (OMOP) with their frequency and proportion

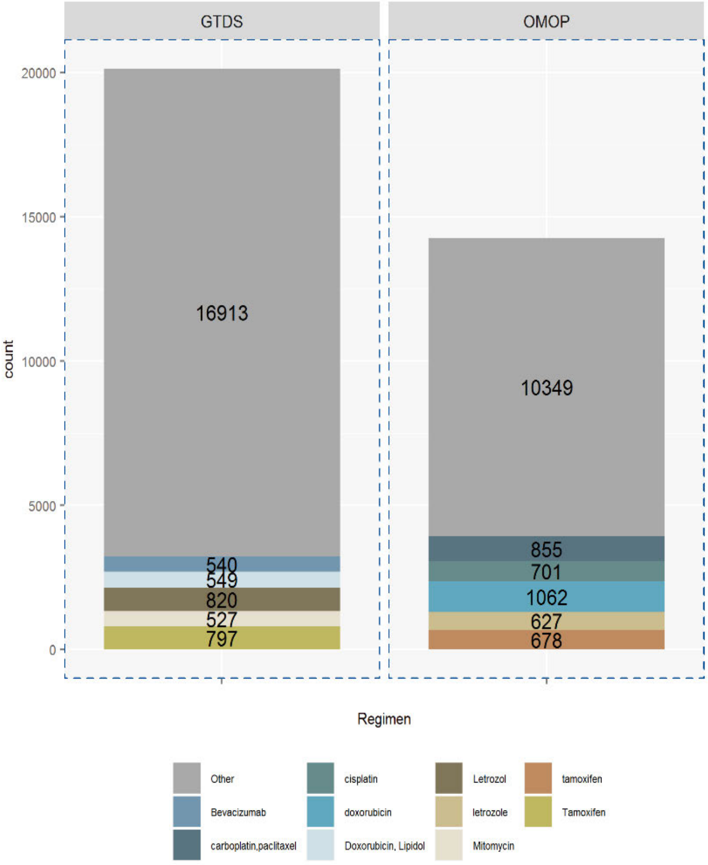


Figure 20: Implemented Treatment Regimen in Source GTDS/OMOP

In the source system, the treatment regimens, letrozole (n=820) and tamoxifen (n=797), which are used in the treatment of breast tumours, are the most common. Through the *OncoRegimenFinder* repository, based on the drug prescription information stored in the source system, ddA (doxorubicin) could be transferred to the target system a total of 1062 times. Data entries describing letrozole and tamoxifen treatment regimens were less frequently submitted to the target system (tamoxifen: n=678, letrozole: n=627), but accounted for a greater proportion in the target system (tamoxifen: 4.8%, letrozole: 4.4%).

The NAACCR vocabulary, which is primarily used in cancer registries in the USA, was used to further specify cancer diagnosis through diagnostic modifiers in the target system. Terminologies, which are used in the source system, such as grading, histology, residual classification, R-status after surgery, were translated to the corresponding NAACCR items. Altogether 17 items of the NAACCR vocabulary could be mapped, so that a total of 424369 records could be mapped to standardized standard concepts, representing NAACCR Ontology in Observation and Measurement domain of CDM. Table11 shows the proportion of implemented NAACCR items depending on their domain. One NAACCR item was not included in this analysis because it belongs to the Metadata domain of the CDM and is not discussed in detail in this paper. In this analysis, 16 NAACCR items with 381184 mapped NAACCR concepts in the target system are analysed.

NAACCR_ITEM	NAACCR_NAME	Prop. [%]
522	Histology	34%
523	Behaviour	34%
400	Primary Site	32%
3844	pathological Grade	19%
1320	Residual Status	15%
910	Pathological Stage Group	14%
880	pT	14%
776	Metastasis	10%
890	pN	7%
970	Clinical Stage Group	6%
960	cM	4%
774	Regional Nodes	3%
950	cN	2%
3855	Her2	2%
900	pM	1%
940	cT	1%

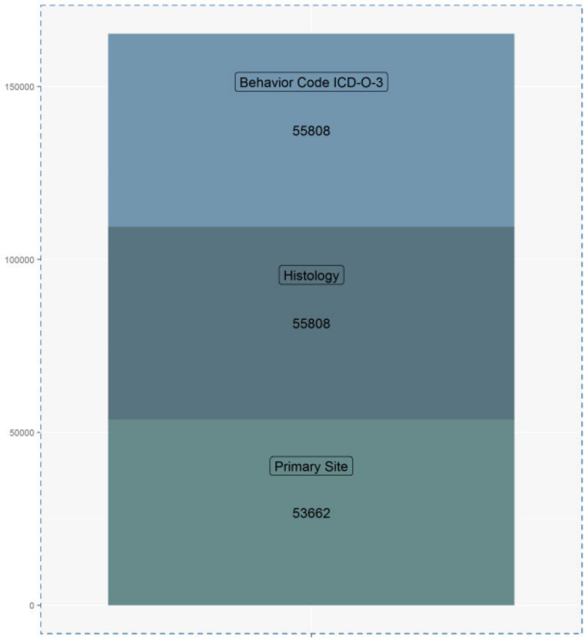
Table 11: Proportion of implemented NAACCR items in the Observation and Measurement domain of CDM

The first three items are assigned to the domain Observation, whereas the proportion of the other items refer to the domain Measurement. The Observation domain contains 55808 data entries for which information on the behaviour and histology of a tumour is available. Data entries that indicate that no information is available on the tumour, depending on their assigned items, also recorded in a standardized manner (e.g X, 0,

..) in source system. However, in the case of histology and primary site, only aprox. 0.11% of data entries are assigned to the NAACCR item of histology or primary site, but coded with the ascription 0, which means that there is no information on the histology of the tumour available (NAACCR Item 522: n= 69, NAACCR Item 400: n=63). Whereas regarding the behaviour of a tumour, a greater detection of data entries could be determined that did not contain any information on the behaviour and were transferred to the Observation domain in the CDM and mapped to standardized standard concept, representing that there is no information available regarding behaviour of a tumour. Of the 55,808 data entries assigned to the behaviour, 3.9% (n= 2190) of the data entries do not contain tumour behaviour information and were marked accordingly in the source system.

Regarding the implemented NAACCR items in the Measurement domain, data records that do not contain any further information on the respective NAACCR item, were also transferred to the CDM in a standardized form. However, only those items that are mandatory documented in the source system are examined in more detail in this results chapter. This includes the declaration of the pathological T-suffix, information of the N- and M-suffix, whereby no distinction is made here as to whether these were assessed clinically or pathologically. In addition, the representation of the R-status after an operation performed in the UKE is examined in more detail, since this is a required information for the ADT/GEKID data set. In the source system surgeries, performed by external hospitals, are also documented. These are filtered out in this analysis. The R-status is used to assess the size of the residual tumour after tumour surgery. This indicates whether a residual tumour is still visible microscopically or macroscopically, or whether the tumour could be completely removed. 12930 data entries for the R-status could be translated to the target system. For 25.6% (n= 3350) of these data entries, no information about the R-status was available at the time of entry into the source system and were therefore coded with an 'X' and mapped to a standardized standard concept, indicating that there is no further information regarding R-Status in target system.

For TNMs, which are marked as relevant for evaluation in the source system, 34165 data entries were available for the pathologically or clinically obtained suffix N and M and 30010 for pathological T suffix. Missing values in the source system were not transferred to the target system. It was observed that 0.5% (n=486) of the recorded pT-suffixes, had no further information about the precision to the size of the primary tumour (~pTX). Regarding clinically or pathologically assessed N- or M-Suffix, there were 1.4% (n=486) data entries, with no further information to specify the suffix (~cNx, ~pNX., ~cMX, ~pMX).



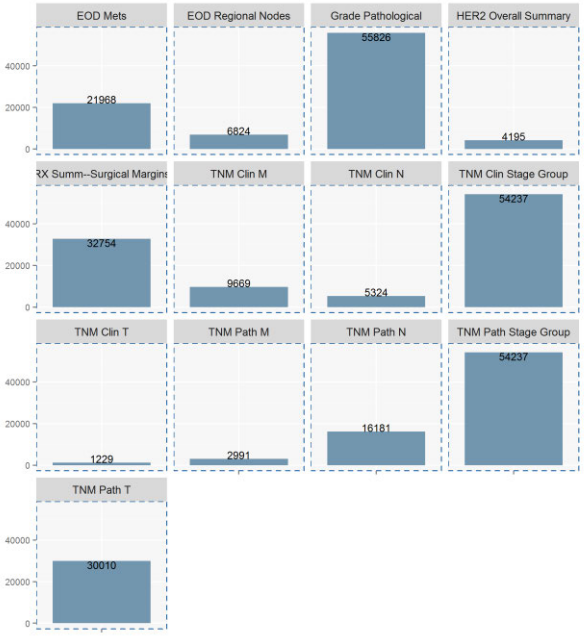


Figure 22: Total amount of implemented NAACCR items in Observation domain

Figure 21: Total amount of implemented NAACCR Items in Measurement domain

4.2 Application of CDM

Regarding the evaluation of the usability of the OMOP model in the clinical cancer registry, two use cases were developed, representing the areas of mapping ADT/GEKID to OMOP and quality assurance. It was originally planned to also represent the research area through an additional use case which should have been achieved through the ATLAS tool of the OHDSI, however, at the end of this work no upgrade of ATLAS for the 6.0 version of the OMOP model was available, so that it was decided to exclude this use case in this work.

4.2.1 Transmission of ADT/GEKID on OMOP CDM v.6.0

To determine the transmission rate from ADT/GEKID data set to OMOP CDM, 20 categories of ADT/GEKID were (see Tab. 12) examined regarding their transferability to OMOP model. Items/Categories which could be implemented by custom extension of the OMOP vocabulary, were assigned the value NULL and not considered in the further analysis. As a result the categories: *Radiation Side Effect*, *Reporting Reason* and *Drug Side Effect* are not considered in this paper. There is one exception in the area of surgeries. During the ETL process the OPS nomenclature has already been officially adapted to the vocabulary of the OMOP model. In this elaboration, the manual mapped OPS ontology is used. This should be changed, as ontologies integrated in the CDM vocabulary are regularly updated and further developed by the OHDSI collaborative. ADT/GEKID items which could be translated to any ontology in the CDM vocabular are assigned the value 1 and is considered as implementable. However, a total of 17 categories of the ADT data set was included in the evaluation. How to query the ADT/GEKID items via OMOP Model can be seen in Appendix 8. Categories of the ADT data set, which can be represented 100% through the OMOP model are: *Course*, *Death*, *Diagnosis*, *Histology*, *Note*, *Other classification*, *Performance*, *Residualstatus*, *Surgeon*, *Drug Therapy*, and *Tumourboard*, *Operation*, *Register data* and *TNM*.

Theme	**not implementable**	**implementable**	**NULL**	**total**	percent [%]
Course	0	5	0	5	**100**
Death	0	3	0	3	**100**
Diagnosis	0	9	0	9	**100**
Histology	0	9	0	9	**100**
Metastasis	1	1	0	2	**50**
Note	0	1	0	1	**100**
Operation	0	3	1	3	**100**
Other classification	0	3	0	3	**100**
patients data	11	3	0	14	**21.4**
Performance	0	1	0	1	**100**
Radiation	0	8	2	8	**100**
Radiation side effects	0	0	3	0	**NA**
register data	0	6	5	6	**100**
Reporting reason	0	0	1	0	**NA**
Residualstatus	0	2	0	2	**100**
Surgeon	0	1	0	1	**100**
Drug side effects	0	0	3	0	**NA**
Drug therapy	0	8	0	8	**100**
TNM	0	10	6	10	**100**
Tumourboard	0	2	0	2	**100**

Table 12: Evaluation regarding the implementability of ADT items subdivided by their categories

In total, 78.5% of the ADT/GEKID base dataset can be represented with the ontologies currently included in the CDM. Since many categories can be completely (=100%) covered by the existing CDM vocabulary, a transfer of the ADT/GEKID to the OMOP ontologies, especially with regard to the promotion of European data integration, makes sense.

4.2.2 Certification

Within the scope of DKG certification, in order to maintain uniform and quality-based patient care in oncological treatment, the provision of treatment data, by means of KPIs defined by the DKG, is required. Depending on the tumour entity, the number and scope of the defined KPIs varies. In this elaboration, the urinary bladder submodule, which is part of the uro-oncology centre, was studied in terms of determining these KPIs using the OMOP model. In order to adequately query the data from the CDM, the characteristics in the source system that are necessary for certification queries were included as custom vocabulary (Certification, Classification, Studycohorts) in CDM. The query for the calculation of the KPIs is defined on two levels. In a first step, the baseline cohort was defined, while subsequently this cohort was examined in a second step with regard to the KPIs defined by the DKG (Appendix 9).

Inclusion criteria for the patient cohort were the “primary case” and “recurrence case” definitions for the urinary bladder module according to DKG guidelines for the year 2019. Figure 23 + 24 shows the percentage fulfilment of the KPIs. In the left figure, the KPIs were generated using the OMOP model, whereas the values determined in the right figure come from the source system. At the time of this report, KPI 13 has not yet been implemented in the source system, which is why it is not discussed further in this report.

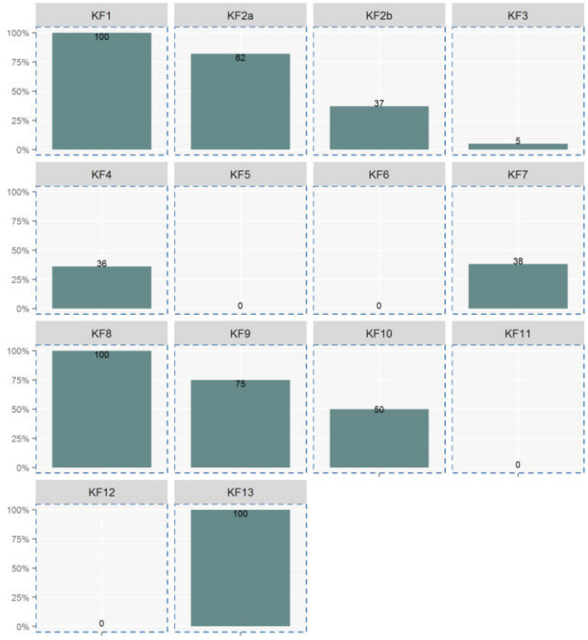


Figure 24: Calculated DKG KPIs for bladder module via OMOP CDM

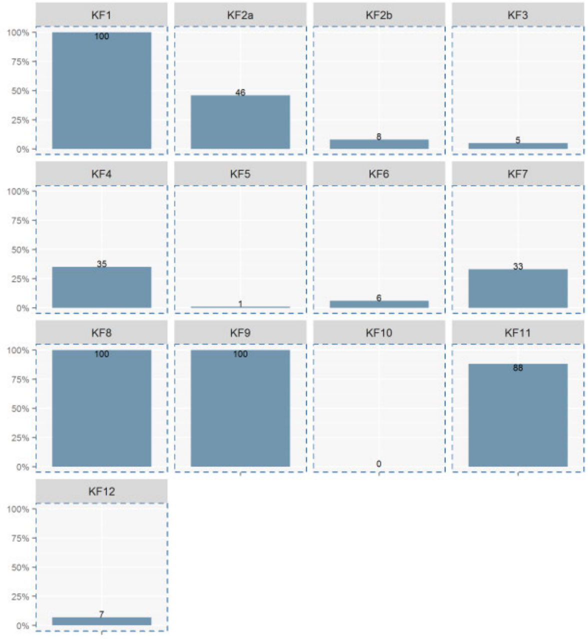


Figure 23: Calculated DKG KPIs for bladder module via medical data warehouse

Through the OMOP model, 129 primary cases and 45 recurrent cases were identified, thus the analysis of KPIs includes a total of 174 centre patients. In the source system, 175 centre patients are represented as patient collective. The source system detects 130 primary cases, whereas only 129 can be detected by the OMOP model. The number of recurrence cases is the same in both patient collectives with 45 recurrence cases. With regard to KPI 1 and 8, both systems meet the DKG's target specifications for the urinary bladder module with 100%. It should be noted, however, that indicator 1 is an absolute target, and the inclusion of 50 primary case patients in the patient population is sufficient for 100% compliance. The OMOP model tends to overestimate with respect to KPIs 2a + b, 4, 7, and 10, whereas it tends to underestimate with respect to KPIs 5, 6, 9, 11, and 12, assuming that the source system represents the correct data. In total, the CDM model seems to slightly underestimate regarding calculation of KPIs when including the arithmetical mean ($\bar{x}$ = -0.5).

5 Discussion

This chapter discusses the implementation of the CDM, especially the mapping of source data to standardized ontologies embedded in CDM Vocabulary and the implementation of the oncology module. In addition, the limitations of this work are highlighted and critically reflected. Hereafter, the two use cases data exchange/analysis through transmission of ADT/GEKID on OMOP CDM and quality assurance through certification will be discussed in detail.

5.1 Implementation of OMOP CDM v.6.0

Overall, the implementation of the OMOP model can be considered as successful. It was possible to transform 1778129 data entries into a standardized concept of the OMOP model in the clinical data area of CDM, which forms the data basis in most of the analyses performed with the OHDSI collaborative tools. With the oncology module there is a total of 1936055 concepts. If considered additionally, that the present elaboration only includes 34 tables, out of 422 possible ones, from the source system into the ETL process; it can be assumed that the value of the standardized concepts will continue to increase with an expansion of the ETL process and the inclusion of more data from the source system. Especially the development of the oncology module and the resulting analysis possibilities for cancer registry data represent a great enrichment for the CCC. The episodic modelling of disease course and therapy approaches in the oncology module extends the possibilities of time series analysis in the UCCH. Furthermore, tools and common queries developed for the Oncology module can be used actively in oncology research or other areas of focus. In addition, by mapping the source data to new ontologies (RxNorm, HemOnc.org, SNOMED-CT), there is now the possibility to exchange data with other institutions through a wider range of ontologies. Thus, with the implementation of new ontologies, it is technically easy possible for UCCH to exchange cancer-related diagnostic data via SNOMED-CT codes, instead of ICD-10-GM, within and across research institution and with the approval of the ethics committee.

Vocabulary

By implementing the OMOP model, new ontologies were added successfully to existing terminologies in the source system. Moreover, the standardization of data representation accelerated or simplified translational and joint research projects within the OHDSI collaborative. Most registry data were mapped into the standardized clinical data area of the OMOP model. Whereas other areas of the OMOP model were not included in the ETL process. For example, the area of standardized health economics of the OMOP model was not considered further in this work, as the source data do not include accounting data and therefore a mapping of the source data to this area of the CDM is not possible. For UCCH, the oncology module and the associated analysis and representation capabilities of cancer data are particularly useful. Currently, the ICD-O-3 is used to link abstracted disease and treatment episodes to lower-level clinical events (Procedure, Observation, Condition, Measurement) of the CDM. Regarding ICD-O-3 linkage, it could be determined that the central curation of integrated ontologies in CDM can also present itself as a disadvantage. At the time of the evaluation for this work, only 3392 distinct ICD-O-3 codes were mapped to the corresponding SNOMED-CT counterpart by the OMOP-developers. If one remains in the domain of condition it is only 1929 distinct codes. Furthermore, ICD-O-3 is only mapped to SNOMED-CT ontology in CDM vocabulary domain. Linking the oncology module to the CDM would be more effective and variable, if developers would map ICD-O-3 vocabulary to other ontologies, which are integrated in vocabulary infrastructure of CDM. This would increase the application possibilities of the oncology module.

In the drug domain, the ATC ontology, which is used as standardized source concepts, is defined within the CDM vocabulary as a classification concept. This allows complex hierarchical queries using the *concept_ancestor* table in the area of standardized vocabularies of the CDM for the proper identification of parent-child and grandparent-grandchild connections. In the area of procedures, the OPS ontology was mapped to the corresponding standardized standard SNOMED-CT concept using the Usagi program, which based on a text algorithm. During this work, the OHDSI vocabulary group included the OPS Ontology in the CDM vocabulary. The mapping was done exclusively by text-matching to the existing CDM vocabulary for the domains *Device* and *Procedure*. Nevertheless, in this paper manually mapped OPS terminology to SNOMED-CT ontology was used. In the future, however, this should be changed, as the application possibilities for the ontologies integrated in CDM vocabulary are better compared to custom vocabulary. Since data exchange within the network is not possible while using custom vocabulary, it can only be used for internal analysis purposes. In addition, OMOP-developers also mapped the OPS codes into the *Device* domain, which was not populated with data entries from the source system in this work. Through the integration of the OPS ontology into the CDM vocabulary, the mapping in the *Device* domain can now be implemented. Furthermore, the drug therapies from the source system were integrated as custom vocabulary in the Procedure domain. During the integration of data entries from the source system, it was noticed that the corresponding codes for the designation of the type of drug therapy (e.g., hormone therapy, chemotherapy, polychemotherapy, ...) were partly outdated in the table directory from the developers of the GTDS and a corresponding documentation about the expression of the item could no longer be found. In addition, the type of drug therapies in the source system are much more granular compared to the possibilities of the SNOMED-CT ontology. Whereby, through the implementation of drug therapy specifications into the CDM, a loss of information, due to the lack of granularity of SNOMED-CT ontology regarding drug therapy specification, is associated (Appendix 10). A lack of granularity, was also noted with respect to tumour irradiation application modes. In the field of radiotherapies, the SNOMED-CT ontology is more granular compared to the source system. An example is the designation of a brachytherapy. SNOMED-CT vocabulary of the CDM, consists of 85 standardized standard concepts regarding brachytherapy specification, whereas only 11 concepts are stored in source system. Different granularity of data representation in the source and target systems was often a problem during manual mapping and needs a constant revision. Especially regarding updates in existing ontologies in the CDM, but also changes and updates in

GTDS, require constant re-evaluations. Changes and updates need to be adjusted to the manual mapping, which comes with time and cost, as this process is not automated.

Oncology Module

Through the implementation of the oncology module and in particular the derivation of source data in time intervals for the assessment of the disease course, the analysis options for time series in the clinical cancer registry are supplemented by time intervals. Time intervals can be used, for example, in the determination of descriptive statistics for the calculation of the average duration of disease phases depending on the tumour entity, overall-survival, progression-free-survival, disease-free-survival or the prediction of disease course depending on the cancer entity. So far, only four concepts are available to assess the disease interval in oncology module (first occurrence, disease remission, disease progression, disease recurrence). However, in the future, these concepts should be supplemented by a concept describing the stable state of the tumour disease. Furthermore, the assessment of disease course in the source system can be displayed in 4 columns in two tables, that have between 10 and 13 expressions. These date events to assess tumour course also include primary metastatic patients who are coded in the source system with a progressive disease measurement point within 4 weeks of diagnosis. The primary metastasized patients were filtered out by means of a function (Appendix 4) but only a subset of the expression options were used to assess disease course. In addition to implementing new concepts in the oncology module for assessing disease course in more detail, the algorithm, which converts date events into time intervals, should be further extended towards the expression options for assessing disease to improve the transfer from date events in the source system to time intervals in the oncology module.

Additionally, it is possible to establish a relationship between the disease episodes and the treatment episodes that took place during that time by using a foreign key in the episode_parent_id column in the *episode* table in the oncology module. This allows a simplified assessment of treatment approaches depending on the disease phase (e.g. targeted therapies vs chemotherapy in disease progression phase). However, the integration of a foreign key to establish relationships between disease and treatment episodes has not yet been implemented at the time of this work and should be done as a next step to extend the application possibilities of the oncology module.

Through the oncology module, HemOnc.org ontology was also integrated into the technical infrastructure of the clinical cancer registry. In the source system, the treatment episodes have so far only been represented in an intern-structured way, which does not limit internal evaluation but makes easy data exchange or common analysis pipelines with other institutions impossible without complex transformation steps (e.g. drug ingredient ≠ treatment regimen (e.g. treatment regimen = combination of drug ingredients)). The HemOnc.org vocabulary represents treatment regimens in a structured form and through the integration into the CDM vocabulary, the HemOnc.org vocabulary is linked to other concepts of the integrated ontologies in the CDM. Thus, it is already possible to query the corresponding RxNorm pedants of the HemOnc.org regimen via the *concept_relationship* table, using a 1:n causality. Furthermore, in forthcoming releases of the Vocabulary Group of the OHDSI collaborative, the HemOnc.org wiki will be further integrated into the CDM vocabulary with additional internal and external relationships. In the future, it should be possible to draw conclusions about the tumour diagnosis from the applied regimens by means of internal relationship. However, general information, which is included in the HemOnc.org wiki, will be also integrated into the CDM vocabulary. HemOnc.org wiki provides information for each treatment regimen that is approved, on the conducted studies that lead to the approval of the treatment regimen. This includes study name, year of enrolment, URL, PMD and journal year. For forthcoming releases of the HemOnc.org ontology embedded in CDM, this information should be queryable.

Limitations

During the implementation of the CDM, the latest version (6.0) of the CDM was integrated into the UCCH test system. From the developers' side, the upgrade of the entire software and analysis tools (R-Packages, Web applications) should have taken place in the third quarter of 2020. Unfortunately, this date was postponed without setting a new date. Therefore, it must currently be said that the use of the software toolchain on top of the OMOP model v.6.0 is considerably limited and an implementation of the 6.0 version is not recommended. Downgrading the ETL process to version 5.3 is possible but costs time. The integration of the CDM in the 6.0 version for local purposes is possible without any problems. But the data basis for research projects within the community is currently still in version 5.3. Before implementing the CDM, the area of use cases should be weighed up and depending on it, it should be decided which CDM version to integrate.

Next to it, it was determined that the mapping of ICD-10-GM to SNOMED-CT was much better than the mapping of ICD-O-3 to SNOMED-CT. This is due to the transformation process that is necessary to generate the ICD-O-3, using regular expression, according to the SEER standard. With automatic transformation processes there is always the danger of an information loss, since a manual check in large datasets for completeness and/or correct generation is only conditionally possible. Here, rule-based mapping can have a positive effect on the loss of information, e.g., by filtering out patients with missing data, before implementing them in target system, and sending them back to the respective clinical coder in the source system for verification.

Furthermore, at the end of this work, the events that provide information on the date of death of a person are only linked to the master data of the patient in the target system. However, it is also possible to record the deaths in a standardised way via the *condition_occurrence* table. This should be implemented in the future in order to integrate the deaths in the OHDSI collaborative applications.

The transformation of the source data to the NAACCR ontology is also not yet complete. On the one hand, the source data is only mapped to 17 items of possible 1881 items of NAACCR ontology and on the other hand, the items are currently not all integrated on value level, which embeds the low-level hierarchy of the NAACCR ontology. To increase the analysis possibility of the NAACCR ontology within the CDM vocabulary, NAACCR values should be implemented in ETL-process in the future. In addition, only a fraction of the source data was mapped into the CDM. The ETL process should therefore be continuously developed, and new source data should be transated to the OMOP CDM. For example, the GTDS also offers the possibility to store genetic mutations of the tumour. Especially with regard to the current focus of cancer research in personal medicine, it would make sense to integrate this data into the CDM as a next step. Furthermore, it should be noted that the linkage of disease and treatment episodes to the low-level clinical events was done for the *condition_occurrence* and *drug_exposures* tables of the CDM. In order to guarantee extensive and variable queries and analyses regarding the disease and treatment episodes, the ETL process for filling the *episode_event* table for linking the low-level clinical events (especially procedures) with the episode abstraction should be further expanded.

5.2 Application of OMOP CDM v.6.0

The application possibilities of the CDM were tested in this work on two use cases, which are beyond the actual scope of the CDM, that was developed for the easy share and analysis of data across different institutions to enable common research projects. The first use case involves the data exchange with the ADT/GEKID via the OMOP model instead of XML. This work examined the suitability of the OMOP CDM, as a single-source information representation model, for easy data exchange with institutions who are using the ADT/GEKID. Transmission of cancer data via a CDM is suitable in principle. Through the ontologies integrated in the OMOP model, 78.5% of the ADT/GEKID base data set can be mapped. This can be considered as a high value. However, it should be noted that the integration of a CDM in the data management area of the research institution is much more complex and requires more data transformation than the integration of an XML schema. Thus, the data of the operational data source in the OMOP CDM must be mapped to a corresponding standard ontology and assigned to the standardization specified by the OMOP developers, through the integrated vocabulary. These complex transformation processes inevitably lead to data loss (e.g., due to failed mapping), which becomes apparent in the cross-institutional data exchange via ADT/GEKID. Thus, these results in an underestimation of the cancer case statistics. Even without the establishment of a medical data warehouse, the transmission of cancer data via XML export of the operational data source is easy possible. Which greatly simplifies the integration of an XML schema at institutions with different technical requirements, compared to the integration of the OMOP CDM. Nevertheless, it should be mentioned that the application possibilities of the data integrated in a CDM are far more complex and diverse than an XML schema offers, especially due to the linking of relationship types. In addition, it should be mentioned that many of the CCCs in Germany use a medical data warehouse in which the integration of the OMOP CDM could be technically possible. Since the ADT/GEKID data set is an accepted data standard and almost every institution in Germany stores cancer data in this data format, an integration of the ADT/GEKID dataset into the CDM vocabulary would make sense. Moreover, the inclusion of the ADT/GEKID in CDM vocabulary would establish relationship to existing ontologies within OMOP. Thus, the diagnoses transmitted to ADT/GEKID by means of the ICD10-GM could be transmitted without much effort by means of SNOMED-CT coding. Since the ADT/GEKID data standard also draws on other ontologies that are used in Germany and have not yet been integrated into the CDM vocabulary, for example Common Terminology Criteria of Adverse Events (CTCAE) to indicate the degree of side effects of an applied therapy, the integration of the ADT/GEKID would be associated with a considerable amount of work, since this would involve the additional integration of further terminologies into the CDM. Nevertheless, the integration of ADT/GEKID could simplify data exchange between research institutions within Germany/Europe and extend the existing infrastructure for data exchange (Lablans et al. 2018: 4f.) and would therefore be a useful step for the support of the European data integration.

The second use case used the example of the bladder module of DKG certification to investigate the extent to which the OMOP CDM can be used to generate certification queries. To ensure common certification queries across institutions with the OMOP CDM, a DKG certification vocabulary should be stored in the CDM vocabulary so that each institution can map its certification data to this standard. For this, there would need to be an overarching implementation of vocabularies that apply to each certification module (primary case, recurrence case, patient case). And on a subordinate level, the implementation of the variables necessary for the determination of the KPIs depending on the organ centre. However, only those variables not covered by the CDM vocabulary should be included, e.g., use of a social-service consultation. KPIs, depending on ontologies which are already integrated in the CDM vocabulary should be linked to the DKG vocabulary. DKG certification queries via the OMOP CDM, build on the same data representation. The comparability of the data would be given by the data harmonization that has taken place. It should be noted, however, that an extensive ETL process, involving many transformation steps, is always accompanied by data loss. Thus, variables for determining the prognostic stage group must be translated to the data standard of the NAACCR vocabulary and then be mapped to the CDM NAACCR standardization before they are read in as modifiers in the *measurement* and *observation* tables. From a European point of view, the implementation of data that specify the cancer therapy or diagnosis in more detail in the CDM is technically very complex, because it uses an ontology, NAACCR, that does not exist in Germany nor Europe. Technically complex data transformations are always accompanied by an increased probability of error accumulation. One solution would be, as mentioned earlier in this paper, the implementation of DKG certification ontology in CDM vocabulary as a German module extension. With a development of a German module for the CDM, time-consuming transformation steps in the context of mapping the source data to the NAACCR ontology could also be omitted, and the probability of data loss in the context of DKG certification could be reduced. Within this module extension it can be also discussed, if European terminologies to specify cancer modifiers are also integrated in the CDM Vocabulary, with the purpose to close the gap of European data integration via the OMOP CDM. In this work, parts of the OPS ontology (primarily those OPS codes that occur frequently in the source system) were mapped to the SNOMED-CT standard. The OPS ontology in the CDM vocabulary was not used because it was added to the CDM vocabulary at a later stage of this work. Therefore, the complete OPS ontology could not be used for the determination of the KPIs, but only those were included in the analysis that occur particularly frequently in the source system. Consequently, some KPIs containing procedures (e.g., KPIs 7, 8, 9, 11, 12) have deviations comparing source and target system. For future work, it is important to investigate whether the OPS ontology added to the CDM vocabulary leads to fewer deviations than the OPS codes mapped manually in the context of this work.

6 Conclusion

Through the implementation of the CDM, the UCCH has become part of the OHDSI collaborative. This enables the institution not only to share data more easily within the community, but also to join collaborative research projects. In addition, due to the successful implementation of the CDM, UCCH has access to a number of analysis tools, including an extensive R-packages library for conducting and analysing observational studies, as well as web applications (e.g., Atlas) that researchers can use to perform real time analysis. Development and maintenance of analysis tools are performed externally by OMOP developers, saving time and costs for the company, that has integrated CDM into their infrastructure. Moreover, centrally stored queries (use case two) can be executed decentral, which saves considerable time and costs but also reduces the probability of mathematical errors, e.g., when conducting a multicentred observational study.

Furthermore, by implementing the CDM, existing ontologies in the source system could be extended with additional ones. This expands UCCH's research capabilities, as the additional ontologies integrated (RxNorm, SNOMED-CT, NAACCR, HemOnc.org, ICD-O-3), facilitate cohort definitions across institutions, regardless of whether they use the OMOP CDM. Thus, it is now possible to transmit administered drugs by means of the RxNorm instead of the ATC code. The USA, for instance, refers to the RxNorm standard when coding drugs. Data exchange between UCCH and an institution in the USA for research purposes, regarding the administration of drugs, is now easily possible. However, it has to be said that the integration of cancer data in the European context into CDM, due to the lack of integration of European ontologies for the precision of cancer diagnosis, is hampered by the need for extensive data transformation steps. Although NAACCR offers a uniform definition for the structured collection of cancer data, this is only used in the USA and Canada. In addition, the NAACCR vocabulary has not yet been mapped to any other standard form in the field of standardized standard vocabularies of the CDM (except SNOMED-CT ontology), which further complicates the integration, especially in the European context and limits analysis possibilities of the NAACCR ontology within the CDMs. Therefore, from a European point of view, the transfer of ADT/GEKID data into CDM Vocabulary could be an important step for European data integration and would close the gap of European cancer representation in OMOP CDM.

Through the integration of standardised concepts in the CDM, homogeneous patient collectives can be formed and examined across institutions, which represents a great benefit for the UCCH, especially in view of the increasingly important approach of personalised medicine. All in all, the implementation of the OMOP CDM in the data infrastructure of the UCCH can be seen as a starting point. It can be used as a solid data basis for cohort characterization, identification of treatment pathways, comparative effectiveness research or medical product safety surveillance for which the CDM was mainly developed. Nevertheless, the OMOP CDM is incredibly variable and can be applied beyond the initial application scenarios, as demonstrated in this work. Especially the integrability of oncological data in the European context should be advanced in the next years. To ensure this, the development of certain extensions could be considered, such as it was briefly described in chapter 5.2 for instance the development of a German module for which it is possible to perform cross-institutional certification queries within the framework of DKG certification, which would ensure comparability of results. Moreover, the integration of the ADT/GEKID into the CDM vocabulary from a European perspective makes sense in order to provide a solid data basis for joint European research projects within the community of OHDSI.

Literature

Arbeitsgemeinschaft Deutscher Tumorzentren (ADT) (2020): Einheitlicher onkologischer Basisdatensatz
URL: https://www.adt-netzwerk.de/einheitlicher_onkologischer_basisdatensatz/basisdatensatz/allgemein/ (access date: 19.10.20)

Belenkaya, R., Gurley, M. J., Golozar, A., Dymshyts, D., Miller, R. T., Wiliams, A. E. et. al (2021): Extending the OMOP Common Data Model and Standardized Vocabularies to Supoort Observational Cancer Research. *JCOC Clinical Cancer Informatics 5. 12-20.*
DOI: 10.1200/CCI.20.00079

Bohland, M. R., Shahn, Z., Hripcsak, G., Tatonetti, N. P. (2015): Birth month affects lifetime disease risk: a phenome-wide method. *Journal of the American Medical Informatics Association 22(2015). 1042-1053.*
DOI: 10.1093/jamia/ocv046

College of American Pathologists (2021): Cancer Protocol Templates.
URL: https://www.cap.org/protocols-and-guidelines/cancer-reporting-tools/cancer-protocol-templates (access date: 25.01.2021)

Damm, R. (2011): Personalisierte Medizin und Patientenrechte. Medizinische Optionen und medizinrechtliche Bewertung. *MedR.* 29: 7-17.
DOI: 10.1007/s00350-010-2826-7

Dennay, M. J., Long, D. M., Armistead, M. G, Anderson, J. L., Conway, B. N. (2016): Validating the extract, transform, load process used to populate a large clinical research database. *International Journal of medical informatics 94. 271-274.*
DOI: 10.1016/j.ijmedinf.2016.07.009

Druker, B. J., Talpaz, M., Resta, D.J, Peng, B., Buchdunger, E., Ford, J. M. (2001): Efficacy and safety of a specific inhibitor of the BCR-ABL Tyrosine Kinase in chronic myeloid leukemia. *The New England Journal of Medicine 344(14). 1031-1037.*
DOI: 10.1056/NEJM200104053441401

Dudeck, J. Wagner, G., Grundmann, E., Hermanek P., Wächter, W., Altmann, U. (1999): Basisdokumentation für Tumorkranke (5. Auflage).
URL: http://www.med.uni-giessen.de/akkk/gtds/grafisch/doku/bd5f.htm (access date: 25.01.2021)

Eggermont, A. M. M., Apolone, G., Baumann, M., Caldas, C., Celis, J. E., de Lorenzo, F. et al. (2019): Cancer Core Europe: A translational research infrastructure for a European mission on cancer. *Molecular Oncology 13. 521-527.* DOI: 10.1002/1878-0261.12447

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L. H. et al. (2000). International classification of diseases for oncology. 3rd edition. World Health Organization (ed.).
URL: https://apps.who.int/iris/handle/10665/42344 (access date: 17.12.2020)

Garza, M., Del Fiol, G., Tenenbaum, J.,Walden, A., Zozus, M. N. (2016): Evaluating common data models for use with a longitudinal community registry. *Journal of Biomedical Informatics 64. 333-341.*
DOI: 10.1016/j.jbi.2016.10.016

German National Cohort Consortium (2014): The German National Cohort: aims, study design and organization. *European Journal of Epidemiology 29(5): 371-382.*

Gruber, T. R. (1993): A translation approach to portable ontology specifications. *Knowledge Acquisition. 5:2. 199-220.*
DOI: 10.1006/knac.1993.1008

Haendel, A., Chute, G. C., Robinson, P. N. (2018): Classification, Ontology, and Precision Medicine. *The New England Journal of Medicine. 379:15, 1452-1462.*
DOI: 10.1056/NEJMra1615014

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B. et al. (1990): Linkage on Early-onset Familial Breast Cancer to Chromosome 17q21. *Science 250(4988). 1684-1689.*

Kimball, R., Caserta, J. (2004): The Data Warehouse ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data. Indianopolis: Wiley

Klein, M., Fensel, D., Harmelen, F.V., & Horrocks, I. (2001): The relation between ontologies and XML schemas.

Kulmanov, M., Smaili, F. Z., Gao, X. et al. (2020): Machine learning with biomedical ontologies.
URL: https://www.biorxiv.org/content/10.1101/2020.05.07.082164v1.full.pdf (access date: 14.12.2020)

Lablans, M., Schmidt, E. E., Ückert, F. (2018): An Architecture for Translational Cancer Research As Exemplified by the German Cancer Consortium. *JCO Clinical Cancer Informatics 2. 1-8.*
DOI: 10.1200/CCI.17.00062

Locker, G. Y., Hamilton, S., Harris, J., Jessup, J. M., Kemeny, N., Macdonald, J. S. et al. (2006): ASCO 2006 Update of Recommendations for the Use of Tumor Markers in Gastrointestinal Cancer. *Journal of Clinical Oncology 24(33). 5313-5327.*
DOI: 10.1200/JCO.2006.08.2644

March, S., Hevner, A. (2007): Integrated Decision Support Systems. A data Warehousing Perspective. *Decision Support Systems 43. 1031-1043.*
DOI: 10.1016/j.dss.2005.05.029

National Cancer Institute (2021): biomarker.
URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker (access date: 02.02.2021)

National Institutes of Health (2020): National Library of Medicine.
URL: https://www.nih.gov/about-nih/what-we-do/nih-almanac/national-library-medicine-nlm (access date: 16.12.2020)

National Institutes of Health (2020): RxNorm Overview.
URL: https://www.nlm.nih.gov/research/umls/rxnorm/overview.html (access date: 17.12.2020)

Netzwerk Onkologischer Spitzenzentren (2020): Das Netzwerk.
URL: http://www.ccc-netzwerk.de/das-netzwerk.html (access date 19.10.2020)

Northern American Assosciation of Central Cancer Registries (2020): Standards for Cancer Registries Volume II. Data Standards. Data Standards and Dictionary. 21 edition.
URL: http://datadictionary.naaccr.org/default.aspx?c=1&Version=21 (access date: 16.12.2020)

Medizinische Uni Giessen (2020): GTDS kurz.
URL: http://www.med.uni-giessen.de/akkk/gtds/gtdskurz.htm (access date: 21.10.2020)

Medizinische Uni Giessen (2008). Kurzanleitung GTDS-Dokumentation.
URL: https://www.med.uni-giessen.de/akkk/gtds/grafisch/doku/dokanleitung.htm (access date: 04.02.2021)

Mehnert, A., Brähler, E., Faller, H., Härter, M., Keller, M., Schulz, H. et al. (2014): Four-Week Prevalence of Mental Disorders in Patients With Cancer Across Major Tumor Entities. *Journal of Clinical Oncology 32(31). 3540-3546.*

Murthy, R.K, Loi, S., Okines, A., Paplomata, E., Hailton, E., Hurvitz, S. A. et al. (2019): Tucatinib, Trastuzumab, and Capecitabine for HER2-Positive Metastatic Breast Cancer. *The New England Journal of Medicine 382(7). 597-609.*
DOI: 10.1056/NEJMoa1914609

Observational Health Data Sciences and Informatics (OHDSI) (2021): HADES – Health-Analytics Data-To-Evidence Suite

URL: https://ohdsi.github.io/Hades/ (access date: 15.03.2021)

Observational Health Data Sciences and Informatics (OHDSI) (2020): OncoRegimenFinder.
URL: https://github.com/OHDSI/OncologyWG/tree/master/OncoRegimenFinder (access date: 12.12.2020)

Observational Health Data Sciences and Informatics (OHDSI) (2020): The Book of OHDSI.
URL: https://ohdsi.github.io/TheBookOfOhdsi/TheBookOfOhdsi.pdf (access date: 25.11.2020)

Observational Health Data Sciences and Informatics (OHDSI) (2019): About OHDSI.
URL: https://www.ohdsi.org/wp-content/uploads/2019/11/OHDSI_1_Pager_v2.pdf (access date: 30.03.2021)

Observational Health Data Sciences and Informatics (OHDSI) (2018): OHDSI Oncology CDM Extension Proposal. Detailed Proposal.
URL:
https://github.com/OHDSI/OncologyWG/blob/master/Oncology.CDM.Proposal.2018-11-15.pdf (access date: 30.10.2020)

Onkozert (2020): Erhebungsbogen. Datenblatt. FAQ.
URL: https://www.onkozert.de/uro/ (access date: 30.10.2020)

Overhage, J. M., Ryan, P. B., Hartzema, A. G, Stang, P. E. (2012): Validation of a common data model for active safety surveillance research *Journal oft he American Medical Informatics Association. 19 (2012). 54-60*
DOI: 10.1136/amiajnl-2011-000376

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M. et al. (2004): A Multigene Assay To Predict Recurrance Of Tamoxifen-Treated, Node-Negative Breast Cancer. *The New England Journal of Medicine 351(27). 2817-2826.*

Perrot, A., Facon, T., Plesner, T., Usmani, S., Kumar, S., Bahlis, N. et al. (2021): Health-Related Quality of Life in Transplant-Ineligible Patients With Newly Diagnosed Multiple Myeloma: Findings From the Phase III MAIA Trial. *Journal of Clinical Oncology 39(3). 227-237.*
DOI: 10.1200/JCO.20.01370

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A. et al. (2001): Use of chemotherapy plus monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *The New England Journal Of Medicine 344(11). 783-792.*

Smith, B., Klagges, B. (2008): Philosophy and Biomedical Information Systems. In Munn, K., Smith, B. (ed.), *Applied Ontology. An introduction.* (Volume 9, p.21-38). Ontos Verlag.

SNOMED-CT (2020): SNOMED CT Concept Model. In SNOMED (ed.), *SNOMED CT Starter Guide.*
URL:
https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model (access date: 17.12.2020)

SNOMED-CT (2021): SNOMED International SNOMED CT Browser.
URL: https://browser.ihtsdotools.org/? (access date: 06.04.2021)

Staniszewska, S., Haywood, K. L., Brett, J., Tutton, L. (2012): Patient and Public Involvement in Patient-Reported Outcome Measures. Evolution Not Revolution. *The Patient - Patient-Centered Outcomes Research 5(2). 79.87.*

Statistisches Bundesamt (Destatis) (2021): Todesursachen.
URL: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Todesursachen/_inhalt.html (access date: 07.02.2021)

Vassiliadis, P., Simitsis, A; Skiadopoulos, S. (2002): Conceptual modeling for ETL processes. In *Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP,pp. 14–21*. New York, USA.

Voss, E. A., Boyce, R. D., van der Lei, J., Rijnbeek, P. R., Schuemie, M. J. (2016): Accuracy of an automated knowledge base for identifying drug adverse reactions. *Journal of Biomedical Informatics 66 (2017). 72-81.*
DOI: 10.1016/j.jbi.2016.12.005

Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M. J. (2015): Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal oft he American Medical Informatics Association 22(2015). 553-564.*
DOI: 10.1093/jamia/ocu023

Warner, J. L., Cowan, A. J., Hall, E. C., Yang, C. (2015): HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals. *Journal of Oncology Practice. 11 (3). 336-350.*
DOI:10.1200/JOP.2014.001511

Warner, J.L., Dymshyts, D., Reich, C. G., Gurley, M. J., Hochheiser, H., Moldwin, Z. H. et. Al (2019): HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. *Journal of Biomedical Informatics. 96 (2019). 1-7.*
DOI: 10.1016/j.jbi.2019.103239

Wu, L., Qu, X. (2015): Cancer biomarker detection: Recent achievements and challenges. *Chemical Society Reviews 44. 2963-2997.*
DOI: 10.1039/c4cs00370e

Appendix

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

Declaration of Authorship

Hereby, I declare that I have composed the presented paper independently on my own and without any other resources than the ones indicated. All direct or indirect sources used are acknowledged as references.

This paper has neither been previously submitted to another authority nor has it been published yet.

Hamburg, 07.04.2021

Jasmin Carus