OPEN ACCESS

Research Article

The Clinical Quality Language as a tool to support data analysis in German clinical cancer registries

Die Clinical Quality Language als Werkzeug zur Unterstützung der Datenanalyse in deutschen klinischen Krebsregistern

Kolja Blohm[1]
David Korfkamp[1]
Joachim Hübner[2]
Florian Oesterling[3]
Stefanie Schulze[3]
Andreas Hein[1]

1 OFFIS – Institute for Information Technology, Oldenburg, Germany

2 Agency for Clinical Cancer Data of Lower Saxony, Oldenburg, Germany

3 Cancer Registry of North Rhine-Westphalia, Bochum, Germany

Abstract

In this paper we evaluate the suitability of the Clinical Query Language (CQL) for supporting various data analysis tasks in German cancer registries. CQL is a domain-specific query language for clinical data. It is particularly used in the United States, for example to define cohorts and to calculate electronic Clinical Quality Measures. We developed a prototype to execute CQL queries on clinical cancer registry data. For this purpose, we created a unified CQL data model for the analysis of clinical cancer registry data. This model is based on the datasets of the cancer registry of North Rhine-Westphalia and the Agency for Clinical Cancer Data of Lower Saxony, as well as the Oncological Basis Dataset. For the evaluation, we applied CQL to typical questions from the areas of guideline-based quality indicators, data requests from external researchers, plausibility checks, and routine reporting, and compared the results with well-established evaluation methods. We were able to show that CQL is capable of representing the complex criteria and temporal relationships that are often relevant for the analysis of data from clinical cancer registries. We see CQL as a promising method to support cancer registries in the definition of patient cohorts for various internal and external analyses and believe that the use of CQL in combination with a standardized data model can make a significant contribution to the standardization of analyses in cancer registries.

Keywords: Clinical Quality Language, CQL, cancer registries

Zusammenfassung

In dieser Arbeit evaluieren wir die Eignung der Clinical Query Language (CQL) zur Unterstützung verschiedener Datenanalyseaufgaben aus der deutschen Krebsregistrierung. CQL ist eine domänenspezifische Anfragesprache für klinische Daten. Sie wird insbesondere in den USA eingesetzt, zum Beispiel zur Definition von Kohorten und zur Berechnung von electronic Clinical Quality Measures. Wir haben einen Prototyp entwickelt, mit dem CQL-Abfragen auf klinischen Krebsregisterdaten ausgeführt werden können. Hierzu haben wir auf Grundlage der Datensätze des Landeskrebsregisters NRW und der klinischen Landesauswertungsstelle Niedersachsen (KLast) sowie des einheitlichen onkologischen Basisdatensatzes ein vereinheitlichtes CQL-Datenmodell für die Auswertung von klinischen Krebsregisterdaten erstellt. Zur Evaluation haben wir typische Fragestellungen aus den Bereichen Leitlinien-basierte Qualitätsindikatoren, externe Datennutzung, Plausibilitätsprüfungen und Routineberichterstattung mit CQL umgesetzt und die Ergebnisse mit etablierten Auswertungsmethoden verglichen. Wir konnten zeigen, dass CQL in der Lage ist, die komplexen Kriterien und zeitlichen Beziehungen abzubilden, die bei der Analyse von klinischen Krebsregisterdaten häufig relevant sind. Wir sehen CQL als eine vielversprechende Methode zur Unterstützung der Krebsregister bei der Definition von Kohorten für verschiedene interne und externe Analysen und sind der

Ansicht, dass die Verwendung von CQL in Verbindung mit einem standardisierten Datenmodell einen wesentlichen Beitrag zur Standardisierung von Analysen in der Krebsregistrierung leisten kann.

Schlüsselwörter: Clinical Quality Language, CQL, Krebsregister

1 Introduction

Cancer registries in Germany are responsible for collecting, processing, and analyzing data from cancer patients. In recent years, the focus has been expanded from an epidemiological to a clinical perspective. Clinical cancer registries collect not only epidemiological data such as sex, date of birth, and diagnosis, but also data on the course of the disease and the treatments performed on individual patients, such as surgeries, systemic, or radiotherapies.

The ultimate purpose of cancer registration is the improvement of oncological care. Registry data are used, for example, to calculate routine indicators for reporting, to assess the compliance with treatment guidelines by means of quality indicators (QI) and to support population-based research. They are used by external researchers and registry analysts, both to answer research questions and to ensure data quality. An important aspect here is the definition of cohorts. To do this, it is necessary to filter the available data based on various criteria. This includes both simple filtering criteria, such as a patient's sex or diagnosis, as well as more complex conditions, such as temporal sequences between different treatments. Defining the criteria and subsequently searching the data based on the definition is a challenge. An example is the calculation of QI 8 from the German clinical practice guideline for breast cancer [1]. This indicator evaluates the frequency of radiotherapies in breast cancer patients after breast conserving therapy (BCT). It is calculated by dividing the number of patients who received radiotherapy after BCT by the total number of breast cancer patients who underwent BCT. The objective is to achieve an adequately high rate of radiotherapy following BCT. According to the German Cancer Society, the percentage should be ≥90% [2].

One way to define and execute such queries on clinical data is by the use of Clinical Quality Language (CQL). CQL is a domain-specific language developed by the Health Level Seven International Organization (HL7) that allows for the precise definition of cohorts in a human-readable format. CQL provides various constructs for formulating queries on temporal data. CQL is particularly used in the United States, for example, to define cohorts and calculate electronic Clinical Quality Measures (eCQMs) [3]. Therefore, the application of CQL also appears promising for clinical cancer registry data.

In this work, we evaluate the suitability of CQL for supporting various data analysis tasks in German cancer registries. To do so, we developed a prototype that enables the execution of CQL queries on clinical cancer registry data.

Our contributions are as follows:

- We describe the CQL data model we created for analyzing clinical cancer registry data in Section 3.1.
- We describe the prototype we developed for defining and executing CQL queries in Section 3.2.
- We demonstrate the applicability of CQL in clinical cancer registries using various data analysis tasks in Section 4.

2 Fundamentals

2.1 Data in German cancer registries

Standardized oncological documentation is an essential aspect of cancer registration, which is handled differently across countries. For instance, Switzerland has a Fast Healthcare Interoperability Resources (FHIR) profile that outlines the format for cancer registrations [4]. In the United States, the North American Association of Central Cancer Registries [5] specifies the reporting channels and exchange formats. In Norway, the reporting channels vary depending on the type of report: Clinical reports are typically submitted through the Kreftregisterets Elektroniske Meldetjeneste (KREMT) reporting portal. Pathology reports, on the other hand, are transmitted using an XML format [6]. In Sweden, cancer cases are reported through a digital reporting form [7].

In Germany, Section 65c paragraph 1 of Book V of the Social Security Code (German: Sozialgesetzbuch V, SGB V) stipulates the use of the unified oncological basis dataset (oBDS, formerly ADT/GEKID basis dataset) for cancer registration [8]. The oBDS is a documentation standard used to report cancer cases to clinical cancer registries. oBDS reports are exchanged in an XML format specified by an XML schema. The schema describes the structure of the oBDS and defines valid value ranges. In addition to diagnosis information, the oBDS also includes data on treatments and disease progression. The oBDS currently summarizes related characteristics in 25 groups, including patient master data, tumor characteristics, therapies applied, adverse effects, and course of the disease. It is currently supplemented by four organ-specific modules (breast, colorectal, skin and prostate cancer). The oBDS is continuously being developed and is currently available in version 3.0.2. [9].

The standardized oncological documentation is an important first step in achieving comparable collection and analysis among cancer registries in different federal states [10]. This creates new analysis options that can support the improvement of cancer care in the long term. To ensure the highest level of data completeness, there

is a legal obligation to report cancer cases. This obligation applies to physicians, dentists and medical institutions involved in the diagnosis or management of cancer. They are required to report on diagnosis, therapy, and course of virtually every malignant tumor disease [10].

Over time, cancer registries typically receive multiple oBDS reports related to a single case of cancer. These reports may come from various sources (such as general practitioners, hospitals, or pathologists) and relate to different medical events (such as diagnosis, treatment, follow-up, or death) ([11], p. 81). To capture the diagnosis, treatment, and disease progression of cancer patients as comprehensively and accurately as possible, the best available information on each case is combined into an analyzable dataset. This dataset is referred to as the *best-of dataset* ([11], p. 81, p. 86).

To determine the most accurate information, there are rules for handling different or conflicting information. Factors such as plausibility, accuracy, source, and timeliness of the information are taken into account. For example, more precise information about localization such as 'upper outer quadrant of the breast' is preferred over less precise information such as 'breast without further specification'. Similarly, pathological information is preferred over clinical information ([11], p. 86-90).

Best-of datasets are important for cancer registries, as they provide a prepared and cleaned basis for analyses and comparisons of data. The German cancer registration manual ([11], p. 86-90) contains various rules for creating best-of datasets. The rules are continuously developed and harmonized by the cancer registries. However, as of now, the rule sets for generating best-of data from the raw reports are not implemented uniformly in the different cancer registries in Germany. Additionally, the best-of data format varies, making it more challenging to conduct comparable analyses between the different cancer registries [12].

2.2 CQL

The Clinical Quality Language (CQL) is a domain-specific query language used, for example, for clinical decision support (CDS) and calculating electronic Clinical Quality Measures (eCQM) [13]. It is designed to express clinical logic that can be read and specified by domain experts. CQL is a standard developed by HL7, which is particularly widespread in the United States. For example, the Centers for Medicare & Medicaid Services (CMS, https://www.cms.gov/), an agency within the US Department of Health and Human Services, and the National Committee for Quality Assurance (NCQA) in the Healthcare Effectiveness Data and Information Set (HEDIS, https://www.ncqa.org/hedis/) use CQL to develop eCQMs [3], [13].

CQL code is organized into libraries in which various elements, such as named expressions, functions, and parameters, can be defined. These elements can be imported into other libraries for reuse. Figure 1 shows the Mamma_QI8 library, which implements a simplified version of the QI 8 for breast cancer [14]. It evaluates the frequency of radiotherapies given to breast cancer patients after BCT. Specifically, it puts the number of breast cancer patients who received radiotherapy after BCT in relation to all breast cancer patients with BCT. In the following, some of the most important CQL constructs are briefly explained using this example.

CQL supports various operators, including logical, comparison and arithmetic operators, as well as operators for working with text, dates, lists, etc. The available data is described in a data model. The data model defines the entities, attributes, associations and data types that can be used within a library. CQL supports several data models, such as the Quality Data Model (QDM) and FHIR, and also allows the definition of new data models. The data model also defines the contexts in which the data can be accessed. The context specifies which data the following instructions refer to. Common examples of a context are *patient* or *practitioner*. The CQL data model referenced here, the *SepamimModel*, for the clinical cancer registry data, is described in more detail in Section 3.1. For the calculation of QI 8, we use the *tumor* context. This evaluates the following statements in the context of a single tumor, so that, for example, the named expression *Is Breast Cancer* evaluates whether the tumor is breast cancer or not. The two expressions *Breast Conserving Therapies* and *Mastectomies* search for surgeries that are breast conserving therapies or mastectomies, respectively, based on their codes according to the Operation Procedure Classification System (OPS) [15]. The surgeries of a tumor are accessed using the retrieve expression *[Surgery]*. In addition to the context, retrieves are the central element in CQL for accessing clinical data. A retrieve returns a list of data and, like the other statements, is executed in the current context. As such, *[Surgery]* returns a list of all surgeries performed on the current tumor. The data that can be accessed with retrieve expressions is defined in the respective data model.

The three previously created expressions are now used to define the criteria for inclusion in the denominator of QI 8. For this purpose, it is checked whether the tumor is a breast cancer for which at least one BCT has been performed and for which no mastectomy has been performed. In addition to the criteria for the denominator, the numerator of QI 8 further restricts the tumors to those that have received at least one radiotherapy within 8 months after BCT. As with the denominator, the criteria for the numerator are first broken down into smaller components. First, the relevant radiotherapies are identified based on the target area of each radiation treatment. Next, the relevant radiotherapies are filtered based on their relationship to the BCT. For this purpose, CQL supports the ability to execute queries on more than one data source. The time interval of 8 months can be expressed using CQL's temporal operators. Finally, the numerator of QI 8 can be defined.

Both the *numerator* and *denominator expressions* return a Boolean value indicating whether the tumor should be included in the respective cohort or not. Additional calcu-

```
library Mamma_QI8

using SepamimModel

context Tumor

define "Is Breast Cancer":
  StartsWith(Tumor.icdCode, 'C50')

define "Breast conserving therapies":
  [Surgery] bct
    where exists bct.ops opsCode
      where StartsWith(opsCode, '5-870')

define "Mastectomies":
  [Surgery] surgery
    where exists surgery.ops opscode
      where StartsWith(opscode, '5-872')
        or StartsWith(opscode, '5-874')
        or StartsWith(opscode, '5-877')

define Denominator:
  "Is Breast Cancer"
    and exists "Breast conserving therapies"
    and not exists "Mastectomies"

define "Relevant Radiotherapies":
  [Radiotherapy] r
    where exists r.radiations radiation
      where StartsWith(radiation.targetRegion, '3.1')
        or StartsWith(radiation.targetRegion, '3.2')

define "Radiotherapies after BCTs":
  from
    "Breast conserving therapies" bet,
    "Relevant Radiotherapies" rad
    where start of rad.period between bet.date and bet.date + 8 months

define Nominator:
  "Denominator"
    and exists "Radiotherapies after BCTs"
```

Figure 1: CQL implementation of a simplified version of the Quality Indicator 8 from the German clinical practice guidelines for breast cancer

lations are necessary for clinical quality measures. In the example, the tumors in the numerator and denominator must be counted and the counts must be divided. Such calculations are not typically performed directly in CQL. Instead, often only the logic for inclusion and exclusion criteria is defined in CQL while further calculation steps take place using external tools. The Quality Measure Implementation Guide (QMIG) describes such an approach, including various types of measures such as calculating proportions and ratios, as well as additional aspects such as stratification [16].

3 Implementation

In this section, we describe the prototype for executing CQL queries on clinical cancer registry data. The structure of the prototype is shown in Figure 2. The prototype uses the best-of datasets, which are already prepared for analysis purposes, as a data basis. These are first transformed into a unified CQL data model. The CQL data model represents the structure of the data available for queries and was created on the basis of the best-of datasets of the cancer registry of North-Rhine-Westphalia (LKR.NRW) and the Agency for Clinical Cancer Data of Lower Saxony (KLast), as well as the oBDS. The CQL data model can be queried via a client-server application. The results of a query are visualized in the application and can be exported for further processing.

3.1 CQL data model

To execute CQL queries on clinical cancer registry data, a suitable CQL data model is required. Although the best-

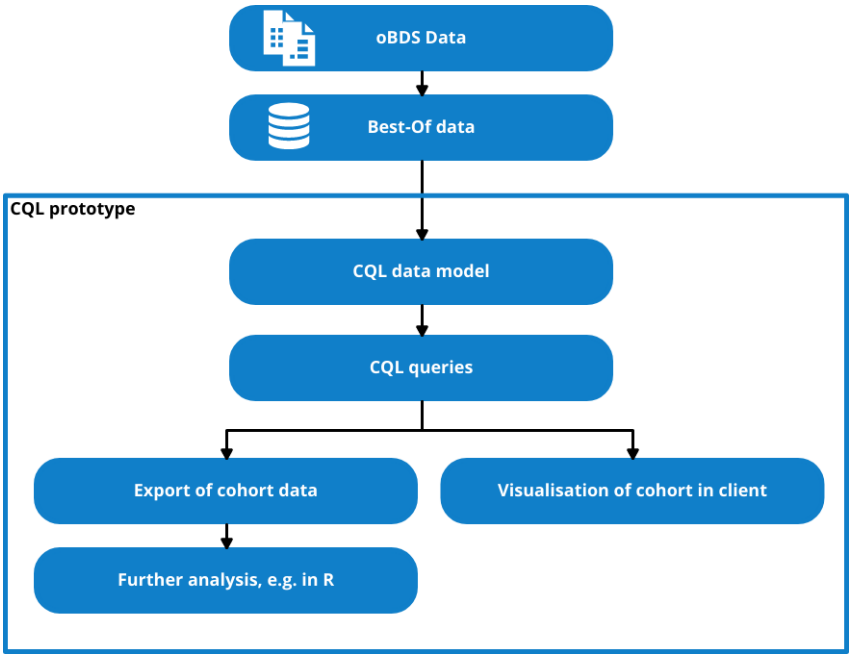

Figure 2: Illustration of the structure of the CQL prototype and the upstream data sources. The oBDS data is prepared in the cancer registries in the form of best-of data. This data is converted into a CQL data model in the prototype and can be queried using CQL queries. The query results are visualized and can be exported for further processing.

of datasets of the LKR.NRW and KLast are based on the oBDS format, they differ in several aspects. This is because cancer registries have implemented individual processing steps to create their best-of datasets. Initially, we compared the best-of datasets of both registries. We found that entities and attributes with the same semantics were sometimes named differently. Additionally, some attributes existed in only one of both models. An example of this are the best-of TNM attributes of the TNM classification for tumor stage [17] at the time of diagnosis. In the KLast, three best-of values are computed: the best-of clinical TNM, the best-of pathological TNM, and a best-of value that is derived from both clinical and pathological data. The combined best-of TNM is not included in the best-of set of the LKR.NRW. Another example is the age of the patient: In the KLast best-of, only the age of the patient at the time of diagnosis is available, while the LKR.NRW best-of contains the date of birth.

Based on these findings, we collaborated with experts from the LKR.NRW and KLast to create a unified CQL data model. We focused on unifying the naming of entities and attributes, as well as calculating missing attributes whenever possible. Attributes or entities that only exist in one dataset and cannot be recalculated are optional in the CQL data model. An overview of the CQL data model is shown in Figure 3. The CQL data model is briefly described below.

In the CQL data model, the individual tumor, not the patient, is the central evaluation unit, as is customary in German cancer registries. This is due to the fact that a patient can have several separate primary tumors. The tumor consists of the best-of data at the time of diagnosis and includes information such as diagnosis, histology, TNM classification, etc., as well as patient-related data such as age at diagnosis and sex. The *tumor* is currently the only entity that can be used as the context of a CQL query.

Information about each treatment is mapped to the *Surgery*, *Systemic Therapy*, and *Radiotherapy* entities. This data can be accessed using retrieve expressions. A *surgery* includes, among others, the intention of the surgery (in particular “curative” or “palliative”), the date of the surgery, a list of OPS codes (the German adaptation of the International Classification of Procedures in Medicine), and a list of side effects. Some important fields for *systemic therapy* are the type of therapy (e.g. “chemotherapy” or “immunotherapy”), the intention of the therapy, side effects, and the start and end date. *Radiotherapy* also includes information about the purpose, side effects and the date of the start and end of treatment. In addition, it contains a list of single radiation series, which make up the radiotherapy as a composite and include detailed information such as the type of radiation, target area, and dose.

The course of the disease is represented by the entities *metastases, histology*, *further classifications*, and *tumor status*. These data can also be accessed through corresponding retrieve expressions. Metastases are described by their location and date of diagnosis. *Histology* includes information on tumor grading and examined lymph nodes, among others. The *Other Classifications* entity provides an option to map clinically relevant classifications that are not otherwise included in the oBDS dataset. The tumor status includes, among others, information on the TNM classification and an assessment of the entire tumor regarding remission, progression, recurrence, etc. This information is distributed across various entities in the oBDS dataset, such as *surgery* or *progression report*, and is processed by us in the *tumor status*. In particular, progression reports are often incomplete or contain only

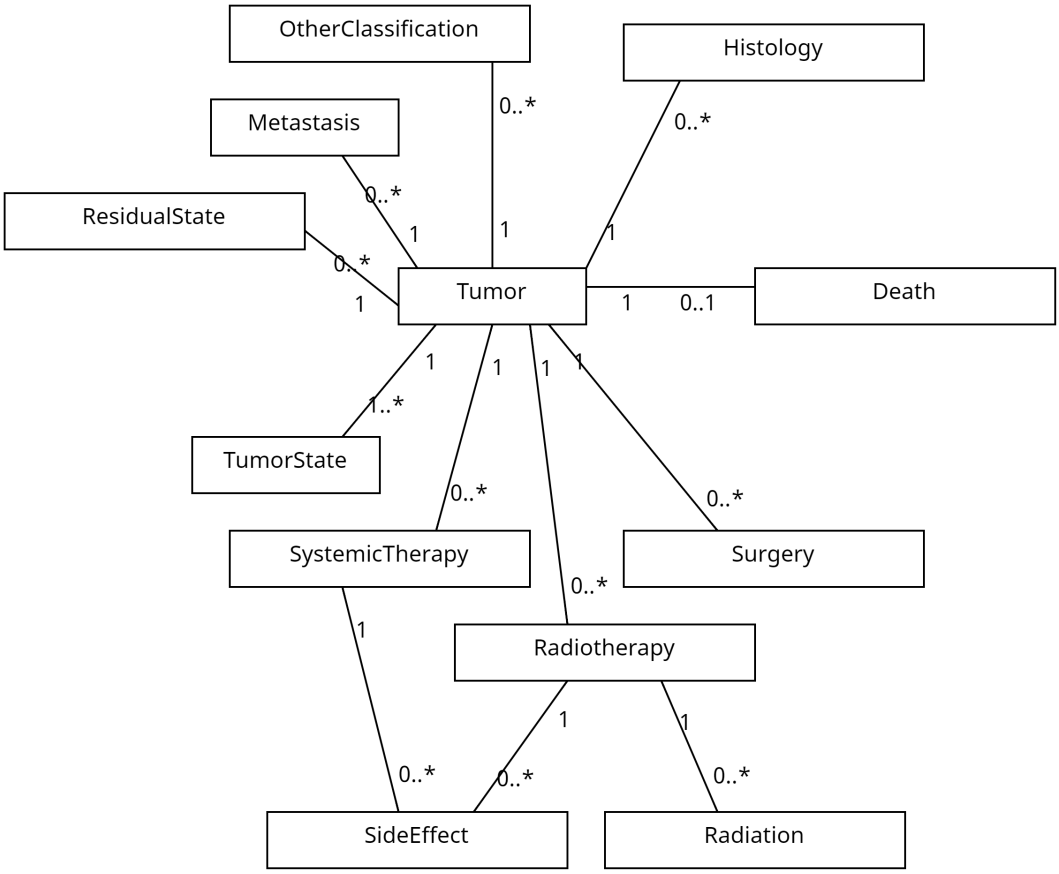

Figure 3: Simplified representation of the CQL data model for clinical cancer registry data. The diagram illustrates the individual entities of the CQL data model and their relationships.

updated information. For example, a TNM T0 value may be reported in an initial report without the other TNM characteristics, and only the TNM M value may be reported in a later report without the T value being repeated. In order to reflect the current status of the tumor as completely as possible, the last reported observation in the *tumor status* is successively carried forward until it is replaced by more recent information. In addition to the actual values, the *tumor status* also has a start and end date, indicating the time period in which the information is valid. Thus, a tumor is continuously described by the information in the *tumor status* from diagnosis to the possible death of the patient. The goal of preparing the data in this way is to simplify queries by eliminating the need to manually search for currently valid values.

The best-of data from the LKR.NRW and KLast are stored in relational databases. To create the CQL data model, they are transformed into a unified format using SQL scripts. This process includes standardizing the names of entities and attributes, calculating missing attributes, and creating the tumor status. Different codings of individual attribute values, such as the coding of UICC stage II as "2" or "II", are currently not being adjusted. Similarly, tumor-specific modules are not yet being taken into account. This is planned for future work.

3.2 Client-server application

To be able to query the CQL data model, we developed a client-server application. The server is a Spring application written in Java and Kotlin. An open source implementation (https://github.com/cqframework/clinical_quality_language) is used to parse and execute CQL queries. The open source implementation allows the definition of custom data sources and data models via appropriate programming interfaces. When the server is started, the data prepared for the CQL data model is read from a relational database into Java objects and stored in memory for later queries. The communication between the client and the server uses a REST interface. In addition to the actual execution of the queries, the results are prepared in the server for display in the client, for example, a Kaplan-Meier survival analysis is calculated.

The web client is a React application written in TypeScript. A screenshot of the web client is shown in Figure 4. CQL queries can be formulated in a CQL editor, which is based on the Monaco code editor (https://microsoft.github.io/monaco-editor) and supports typical features of a code editor, such as syntax highlighting, tooltips with information on data types, etc. Expressions that return a Boolean value can be executed and determine whether or not a tumor is included in the result.

Another important aspect in the development of the application is the visualization of the results of a query. The

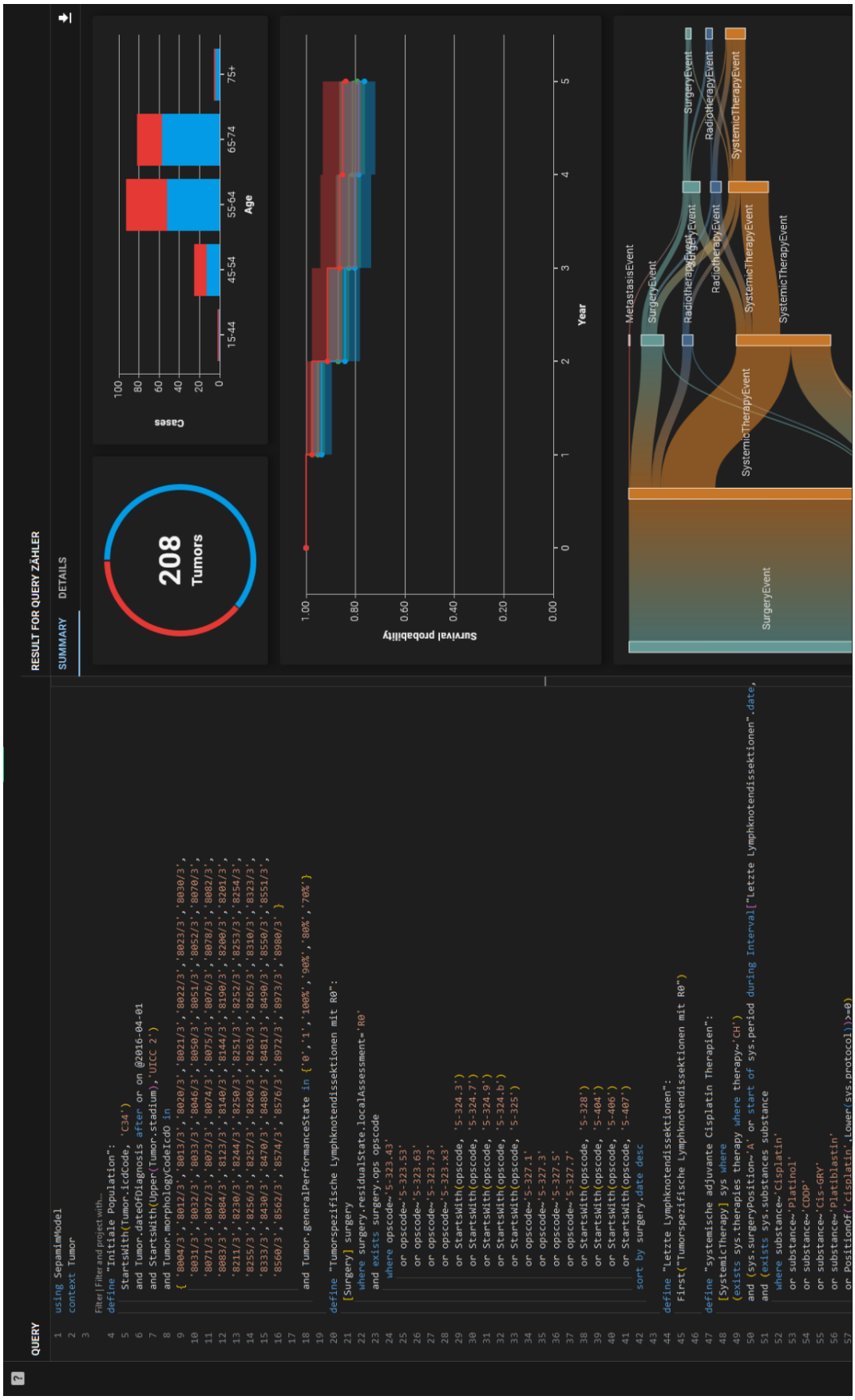

Figure 4: Screenshot of the CQL Web Client. On the left is the code editor, where CQL queries are formulated. Define statements that return a Boolean value can be executed to filter the data. The result of a query is displayed in the right panel. At the top is the sex distribution, and next to it is the age distribution. In the middle is the Kaplan-Meier survival curve of the queried cohort, and below that is the visualization of treatment pathways in a Sankey plot.

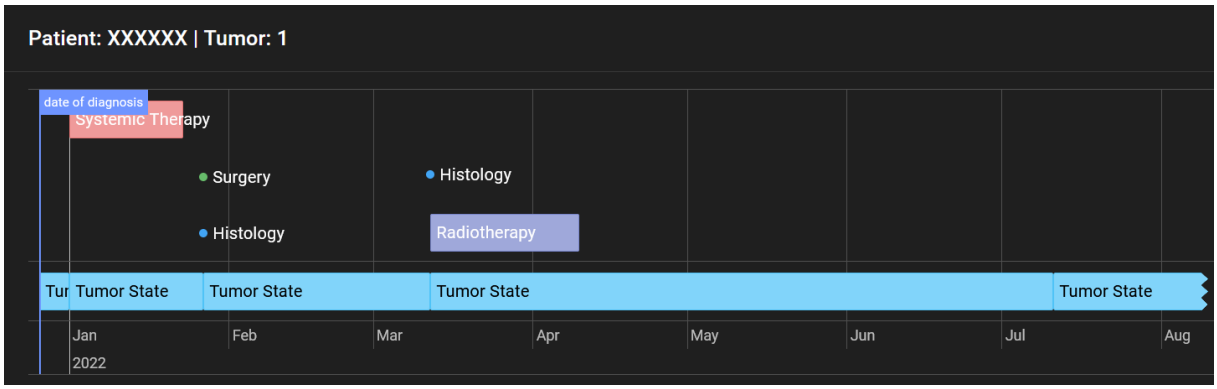

Figure 5: Illustration of a course of treatment. The screenshot shows a section of the patient detail view. Various therapy events of a patient can be seen on a timeline.

prototype provides an overview of some important characteristics of the queried cohort. For example, the age and sex distribution, as well as survival times are visualized as diagrams. In addition, an overview of the treatment pathways is presented in the form of a Sankey plot. The software also provides a detailed view for examining individual tumors, displaying the best-of data available for each tumor. A timeline visualizes the course of treatment (refer to Figure 5), and the properties of each treatment can be displayed.

The results of a query can be exported. In addition to specifying inclusion and exclusion criteria to determine the cohort, a projection of the data can be formulated in CQL. The projected results can then be downloaded as a JSON file.

4 Evaluation

In this section, we evaluate whether CQL is suitable to support clinical cancer registries in performing typical data analysis tasks and whether our CQL data model contains all the relevant information for this purpose. For this, we have selected nine different questions from four categories and implemented them with CQL. These are questions from the LKR.NRW and the KLast, which have already been evaluated with other methods and whose results are compared with the CQL queries. We have queried the required cohorts with CQL and, if necessary, carried out further calculation steps, such as stratification and calculation of measures, with external tools.

4.1 Guideline-based quality indicators

Clinical guidelines are systematically developed statements to support practitioners and patients in making decisions about appropriate health care for specific circumstances. Quality indicators make compliance with selected recommendations measurable by defining rules for calculating related key figures. The QI working group of the platform § 65c (https://plattform65c.de/qualitaetskonferenzen/leitlinien-qualitaetsindikatoren/), an expert panel of the German cancer registries, defines and publishes inclusion and exclusion criteria for the calculation of the QIs on the basis of registry data. We have implemented the QI 8 for breast cancer and the QI 6 for lung cancer using CQL. The QI 8 for breast cancer determines the proportion of patients with invasive breast cancer and BCT who received adjuvant radiotherapies, among all patients with BCT. Quality Indicator 6 for lung cancer determines the proportion of patients with lung cancer who received cisplatin-based chemotherapy after a complete tumor resection, among all patients who underwent a complete tumor resection.

The CQL scripts for both QIs were implemented based on the description of the platform § 65c [14], [18]. For this, the inclusion and exclusion criteria for numerator and denominator were implemented and compared with existing results from the LKR.NRW. Initially, there were differences in the results. After a discussion based on the CQL scripts, we were able to determine that the differences occurred due to different interpretations of the inclusion and exclusion criteria. After correcting the SQL scripts and adjusting to a uniform interpretation of the QIs, both implementations achieved identical results. The initial differences in the results suggest that the informal description of the QIs allows for too much room for interpretation.

4.2 Data requests from external researchers

Cancer registries not only use their data for their own reporting obligations and analyses, but also make them available to external users, particularly in the public health and scientific communities ([11], p. 138). In principle, four types of data use can be distinguished ([11], p. 138): individual patient access, cohort matching, aggregated data, and individual case data. In the case of individual patient access and cohort matching requests, data is requested on the basis of personal identifying attributes such as name and address. For privacy reasons, these data are not available in the cancer registries' data analysis centers and therefore are not included in the best-of datasets. As a result, these forms of external data usage will not be further considered here. Aggregated data are individual case data grouped according to various

characteristics, such as ICD-10 diagnosis code, age at diagnosis, or sex. Aggregate measures, such as rates or case numbers, are calculated for the individual groups. Aggregated data can be requested from cancer registries, but are also published in routine reports such as annual data and activity reports and interactive online reports ([11], p. 138-140). The evaluation of CQL queries for aggregated data is discussed in Section 4.4 using the example of queries for the KLast data reports.

Cancer registries can also provide anonymized individual case data for research upon request. A request must define clear inclusion and exclusion criteria as well as the attributes required for the research question ([11], p. 139-140). For the evaluation, we implemented two requests for external use of data submitted to the KLast using CQL. Both requests were made in the context of a study investigating the impact of the time interval between the diagnosis and the start of tumor specific therapy on the outcome in patients with advanced cancer diagnoses. The cancer diagnoses considered were colorectal cancer and breast cancer.

For each data export, the scientific benefit of the data must be weighed against the risk of patient re-identification. Therefore, in addition to filtering the data based on inclusion and exclusion criteria, it is important to export only the attributes that are needed to answer the specific question, and at the appropriate granularity (level of detail). For example, only selected tumor and treatment attributes were exported. Additionally, instead of the date of diagnosis, only the year of diagnosis was exported, and instead of the date values of the treatments, the exact number of days since the day of diagnosis was exported. In addition, the ICD code of the diagnosis was exported only at the three-digit level. The final determination of inclusion and exclusion criteria, attributes, and granularity was coordinated between the applicant and the KLast staff.

The data filtering and projection into the export format were implemented in CQL for both queries. The results of the CQL queries are identical to the existing SQL queries. However, the prototype exports JSON files while the SQL queries export CSV files.

4.3 Ensuring data quality

The quality of the underlying data is an important prerequisite for meaningful analysis of clinical questions. An important task of cancer registries is therefore to ensure data quality. For this purpose, various plausibility checks exist, of which we have implemented three typical examples from the LKR.NRW in CQL. The results are briefly discussed below.

- **Plausibility check 1:** Clinical events after date of death
  This plausibility check identifies tumors for which treatment events were reported with a date after the date of death. This suggests that either the treatment events or the death date are incorrectly dated. Out of a total of 1,160,547 tumors, there were 176 tumors with surgery after the date of death, 197 tumors with radiotherapy after the date of death, and 1,611 tumors with systemic therapies after the date of death. In this plausibility check, we were able to exactly reproduce the results of the SQL query with CQL.
- **Plausibility check 2:** Change from palliative to curative therapy intention
  This plausibility check examines whether there are tumors for which a change from a palliative to a curative therapy intention has occurred in the course of treatment. Such a change is unlikely and should be investigated further. There were discrepancies between the SQL and CQL queries in this plausibility check. In total, the CQL query found 3,898 tumors and the SQL query found 3,896 tumors. The SQL query found 2 tumors that the CQL query did not find. Conversely, the CQL query found 4 tumors that the SQL query did not find. The discrepancies are due to the fact that when preparing the CQL dataset, we interchanged the start and end dates of treatment events if the start date was after the end date. This created valid time intervals, but also changed the chronological order of some treatment events. This resulted in the plausibility check discrepancies.
- **Plausibility check 3:** Tumor size decrease without therapy
  This plausibility check examines whether there are tumors with an implausible decrease in tumor size. A decrease in tumor size should only occur if there was a therapy between the changes in the TNM T value that led to the decrease. The existing SQL query from the LKR.NRW checks whether there are two TNM events between which the TNM T value has improved without any treatment having taken place. However, the individual TNM events are no longer included in our CQL data model. Instead, they are summarized in the tumor status and further processed as described in Section 3.1. Therefore, a direct mapping of the SQL query in CQL is not possible with our CQL data model. However, we have implemented the query in CQL using the tumor status. Due to the different data modeling and preparation, the results are not directly comparable and vary as expected. In total, the CQL query found 2,370 tumors and the SQL query found 8,568 tumors. The SQL query found 7,048 tumors that the CQL query did not find. Conversely, the CQL query found 850 tumors that the SQL query did not find.
  A manual examination of the differences revealed the following: The additional tumors found with SQL are mainly those for which the TNM-T value improved in the first six months after diagnosis. This improvement is not recorded in the CQL data model because the tumor status in the LKR.NRW during this period is based on the best-of values of the tumor at the time of diagnosis. This means that the tumor events of the first six months are summarized in a single best-of tumor status, resulting in fewer changes in the TNM-T value in the CQL data model. The additional tumors found with CQL can be attributed to the fact that re-

ported TNM-T values remain in the tumor status until they are replaced by more current information. In contrast, the TNM events used in the SQL query may include events without a specified TNM-T value. In such cases, the last valid TNM-T value is used in the CQL data model. It should be considered whether it would be useful to expand the CQL data model to include the original TNM events.

4.4 Aggregated data

Routine reporting by German cancer registries serves to inform on and monitor cancer cases in their respective catchment areas. This includes, for example, the publication of aggregated cancer data in the form of annual reports or interactive data query tools on the Internet. Reporting is based on various characteristics such as age, sex, place of residence, tumor type, tumor stage, histology and therapy. An important tool for routine reporting by the KLast is the interactive online report (https://www.klast-n.de/files/interaktiver-bericht/), which provides a comprehensive overview of the cancer situation in Lower Saxony. For the evaluation, we have implemented the analyses for two pages of the annual report with CQL. The first page (https://www.klast-n.de/files/interaktiver-bericht/#/allg/table/) provides an overview of new cancer cases and lists the number of tumors by diagnosis group, sex, year of diagnosis and place of residence (in or outside of Lower Saxony). The second page (https://www.klast-n.de/files/interaktiver-bericht/#/diag/thera/, entry “Operationen“) presents information on the performed surgeries. The number of tumors is again shown by diagnosis group, sex, year of diagnosis, and place of residence, and additionally by UICC stage and various treatment-specific characteristics such as time between diagnosis and first surgery, intention of first surgery, and local residual status.

We implemented both data filtering and feature definition for stratification in CQL. The data was then exported as JSON files and further processed in R. In R, we grouped the data based on the previously determined stratification features and calculated the necessary measures. The results were compared with the results obtained using the methods used in KLast for creating the interactive online report. This report is currently generated using MDX queries on a data warehouse, followed by processing with a C# tool. The results are consistent with one another.

5 Discussion

We implemented several data analysis tasks from cancer registration using CQL and demonstrated that CQL is a suitable tool to support cancer registries in forming cohorts for various internal and external analyses. We were able to show that CQL is capable of representing the complex criteria and temporal relationships that are often relevant for the analysis of data from clinical cancer registries. In our opinion, CQL queries are generally shorter and more readable than existing SQL queries, which could facilitate communication and validation of queries with experts. Additionally, CQL allows for the reuse of named expressions and the definition of libraries, which increases the maintainability and extensibility of the queries.

We found that using CQL for the standardized formulation of QIs is beneficial. Its formal language might help to avoid different interpretations and inconsistent implementations of the QIs. For example, during the evaluation, the results initially deviated from the reference implementation, with the deviations being due to different interpretations of the QIs. Moreover, we found that deviations were caused by the use of different data models.

The evaluation showed that CQL is suitable for exporting data to external users, particularly for exporting cohorts to external researchers and, as we discuss below, with limitations for exporting aggregated data. A uniform query language could enhance communication and facilitate shared use of queries among different cancer registries and external users. This would be beneficial for the use cases mentioned. For example, efforts are being made to publish comparable values in various data reports. Using CQL on a unified data model to define the queries for these reports could be a possible step to improve the comparability of results. Similarly, CQL could be utilized to define queries for data export requests, particularly when data is needed from multiple cancer registries.

During our evaluation of CQL for data quality assurance queries, we found that it is also appropriate for such tasks. However, we observed discrepancies between the results obtained with CQL and the existing SQL queries. These discrepancies were not caused by CQL, but rather by differences in the data models. Some quality issues that were present in the original data had already been resolved in the CQL data model. For a detailed explanation of the differences, see Section 4.3.

However, while CQL has its advantages, it also has some drawbacks. One of these is that it is less widely used, which means that users may need to familiarize themselves with the syntax and semantics of the language. Additionally, CQL is not as powerful as SQL or R. For example, CQL does not offer the ability to calculate measures such as rates, ratios, or survival analyses, or to stratify data, which are common tasks in the field of clinical cancer registries. Therefore, CQL is not a comprehensive solution for analyzing data from clinical cancer registries. Instead, it is a component of a larger set of tools and standards. CQL is mainly used to define inclusion and exclusion criteria for cohorts. Any additional calculations or analyses with the cohort data must be conducted using external tools. The Quality Measure Implementation Guide [16] outlines one such approach. This guide explains how to calculate different types of measures, including rates and ratios, and covers other aspects such as stratification. It is common practice to separate the definition of data required for analysis from the data processing, as is done when processing SQL

results with R. This separation is advantageous as it reduces the complexity of CQL and makes it easier to use.
In order to execute CQL queries on clinical cancer registry data, an appropriate data model is required. In the context of German cancer registries, a standardized data model for reporting event-related information on cancer cases already exists in the form of the oBDS dataset. However, there is no standardized data model for analyses. Here, cancer registries create individual best-of datasets for analysis. Although the best-of datasets are based on the oBDS format, they differ in various aspects, such as the rules used to create the best-of as well as the naming and the selection of entities and attributes. This makes it difficult to perform comparable analyses between different cancer registries. To address this issue, we created a CQL data model based on the oBDS dataset and the best-of datasets of the LKR.NRW and KLast. This model can serve as a basis for a standardized analysis data model. For this purpose, we have standardized the names of entities and attributes, recalculated some best-of attributes (some of which are only available in one of the two best-of datasets), resolved data issues, and generated new best-of values. For instance, we corrected time intervals where the start date was after the end date and created a best-of from the progression reports in the tumor status entity. The results of some queries were affected by the additional data preparation, as described in Chapter 4.
There are three aspects of data preparation that we have not yet taken into account: standardizing attribute values, calculating best-of data uniformly, as well as incorporating tumor-specific modules. To achieve uniformity in attribute values, the classifications for cancer registration published in the oBDS-RKI dataset [19] could be used as a basis. The oBDS-RKI dataset outlines the structure and content for delivering cancer registry data to the Center for Cancer Registry Data (ZfKD) and has been in place since 2023. The ZfKD is an institution of the Robert Koch Institute that consolidates and evaluates data from the cancer registries of the federal states for the entire country of Germany. Additionally, it would be interesting to investigate whether the oBDS-RKI dataset can provide further insights to create a unified CQL data model. Several working groups of the § 65c platform are currently working on the uniform calculation of best-of data. As part of this work, we only had access to pre-processed best-of data, so we had limited influence on its calculation. The integration of tumor-specific modules is planned for a future version.
We would like to briefly clarify our decision not to adopt FHIR, despite its widespread use and standardization as a data model for clinical data. Our decision was based on our desire to align closely with the oBDS format and the best-of datasets of the cancer registries, both of which are well-established in cancer registration. As a result, we opted to develop our own CQL data model. This approach also gave us the opportunity to customize and expand the model according to our needs.

During our work, it became evident that presenting the results of a CQL query is an important aspect. In addition to traditional representations like age and sex distribution, presenting temporal data proved to be challenging. Currently, we use a Sankey diagram to provide an aggregated overview of the treatment data, which displays the distribution and sequence of treatments for a cohort. The course of treatment of an individual tumor can be displayed as a time series. We believe that visualizing treatment data is an important topic in the field of cancer registration, which requires more attention and research.
We have benefited from a variety of resources within the CQL community. We would like to highlight the comprehensive documentation on CQL and the fact that we were able to draw on various open-source projects in the development of our prototype. Two areas that still have room for improvement are the execution speed of the CQL engine and the functionality of the editor tooling. Although there is an open-source language server for CQL, it currently has limited functionality. Therefore, we implemented the editor tooling for our prototype ourselves. However, there is still room for improvement in our CQL editor. For instance, an auto-complete function would significantly facilitate the writing of CQL queries. Our goal is to further develop the prototype in the future to increase user-friendliness.

6 Conclusion and outlook

In this work, we evaluated the suitability of CQL for implementing various data analysis tasks in German cancer registration. We created a CQL data model based on the oBDS dataset and the best-of datasets of the LKR.NRW and KLast, which can serve as the foundation for a standardized analysis data model. We implemented various questions in the areas of guideline-based quality indicators, data requests from external researchers, plausibility checks, and routine reporting using CQL and compared the results with existing evaluation methods. Our findings demonstrate that CQL is capable of representing the complex criteria and temporal relationships that are often necessary for querying data from clinical cancer registries.
We conclude that CQL is a promising tool to support German cancer registries in the creation of cohorts for various internal and external analyses. We believe that the use of CQL in conjunction with a standardized data model can make a significant contribution to the standardization of analyses in cancer registries. In the United States, CQL is already being used for this purpose in the health sector. In the future, we want to further develop and improve the CQL data model and, in particular, integrate additional cancer registries and the ZfKD into this process.
Furthermore, we plan to expand and optimize our prototype to enhance its user-friendliness. To achieve this, we intend to conduct a user study to evaluate the applicability of CQL for specialist users.

Notes

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Federal Joint Committee German Innovation Fund within the joint research project SePaMiM (grant no. 01VSF20018).

References

1. Leitlinienprogramm Onkologie, editor. S3-Leitlinie Früherkennung, Diagnostik, Therapie und Nachsorge des Mammakarzinoms [Evidence-based Guideline for the Early Detection, Diagnosis, Treatment and Follow-up of Breast Cancer]. AWMF-Registernummer 032-0450L. [cited 2024 Mar 14]. Available from: https://www.leitlinienprogramm-onkologie.de/leitlinien/mammakarzinom/
2. Deutsche Krebsgesellschaft, editor. Kennzahlenauswertung (2023). Jahresbericht der zertifizierten Brustkrebszentren. Auditjahr 2022 / Kennzahlenjahr 2021. Berlin: Deutsche Krebsgesellschaft; 2023 [cited 2024 May 23]. Available from: https://www.krebsgesellschaft.de/jahresberichte.html
3. Brandt PS, Kiefer RC, Pacheco JA, Adekkanattu P, Sholle ET, Ahmad FS, Xu J, Xu Z, Ancker JS, Wang F, Luo Y, Jiang G, Pathak J, Rasmussen LV. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. Learn Health Syst. 2020 Oct;4(4):e10233. DOI: 10.1002/lrh2.10233
4. Bundesamt für Gesundheit (BAG). CH CRL (R4) FHIR Bundle. [cited 2024 Mar 14]. Available from: http://fhir.ch/ig/ch-crl/StructureDefinition-ch-crl-bundle.html
5. North American Association of Central Cancer Registries. Central Registry Standards. XML Data Exchange Standard. [cited 2024 Mar 14]. Available from: https://www.naaccr.org/xml-data-exchange-standard/
6. Cancer Registry of Norway. Reporting. [cited 2024 Mar 14]. Available from: https://www.kreftregisteret.no/en/The-Registries/Reporting/
7. Socialstyrelsen. Lämna uppgifter till cancerregistret [Submit information to the cancer registry]. [cited 2024 Mar 14]. Available from: https://www.socialstyrelsen.se/statistik-och-data/register/lamna-uppgifter-till-register/cancerregistret/
8. Sozialgesetzbuch (SGB) Fünftes Buch (V). Gesetzliche Krankenversicherung. Artikel 1 des Gesetzes v. 20. Dezember 1988, BGBl. I S. 2477.
9. Arbeitsgemeinschaft Deutscher Tumorzentren e.V. Bundeseinheitlicher Onkologischer Basisdatensatz. [cited 2024 Mar 14]. Available from: https://basisdatensatz.de/
10. Epidemiologisches Krebsregister Niedersachsen; Klinisches Krebsregister Niedersachsen; Klinische Landesauswertungsstelle Niedersachsen, editors. Krebs in Niedersachsen – Jahresbericht 2020 mit Datenreport 2017-2018. Dez 2020 [cited 2024 Mar 14]. Available from: https://www.krebsregister-niedersachsen.de/veroeffentlichungen/jahresberichte/
11. Stegmaier C, Hentschel S, Hofstädter F, Katalinic A, Tillack A, Klinhammer-Schalke M. Das Manual der Krebsregistrierung. München: W. Zuckschwerdt Verlag; 2019.
12. Katalinic A, Halber M, Meyer M, Pflüger M, Eberle A, Nennecke A, Kim-Wanner SZ, Hartz T, Weitmann K, Stang A, Justenhoven C, Holleczek B, Piontek D, Wittenberg I, Heßmer A, Kraywinkel K, Spix C, Pritzkuleit R. Population-Based Clinical Cancer Registration in Germany. Cancers (Basel). 2023 Aug 2;15(15):3934. DOI: 10.3390/cancers15153934 2023
13. Clinical Decision Support Workgroup. Clinical Quality Language (CQL). [cited 2024 Mar 14]. Available from: https://cql.hl7.org/
14. Leitlinien Qualitätsindikatoren. Mammakarzinom. QI 8 – Durchgeführte Strahlentherapie nach BET. Platform § 65c; [cited 2024 Mar 14]. Available from: https://plattform65c.atlassian.net/wiki/spaces/LLQI/pages/100600499/QI+8+-+Durchgef+hrte+Strahlentherapie+nach+BET
15. Bundesinstitut für Arzneimittel und Medizinprodukte. Operationen- und Prozedurenschlüssel (OPS). [cited 2024 Mar 14]. Available from: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/OPS-ICHI/OPS/_node.html
16. HL7 International. Quality Measure Implementation Guide (STU5). Quality Measure Implementation Guide homepage. [cited 2024 Mar 14]. Available from: https://build.fhir.org/ig/HL7/cqf-measures/
17. Brierley JD, Gospodarowicz MK, Wittekind C, editors. TNM classification of malignant tumours. 8th edition. John Wiley & Sons; 2017.
18. Leitlinien Qualitätsindikatoren. Mammakarzinom. QI 6 – Indikation zur Sentinel-Lymphknotenbiopsie. Platform § 65c; [cited 2024 Mar 14]. Available from: https://plattform65c.atlassian.net/wiki/spaces/LLQI/pages/100599215/QI+6+-+Indikation+zur+Sentinel-Lymphknotenbiopsie
19. Meisegeier S, Imhoff M, Berg K, Kraywinkel K. Bundesweiter klinischer Krebsregisterdatensatz – Datenschema und Klassifikationen (oBDS_v3.0.0.8a_RKI). Zenodo; 2023. DOI: 10.5281/zenodo.10022040

Corresponding author:
Kolja Blohm
OFFIS e.V., Escherweg 2, 26121 Oldenburg, Germany
kolja.blohm@offis.de

Please cite as
Blohm K, Korfkamp D, Hübner J, Oesterling F, Schulze S, Hein A. The Clinical Quality Language as a tool to support data analysis in German clinical cancer registries. GMS Med Inform Biom Epidemiol. 2024;20:Doc11.
DOI: 10.3205/mibe000267, URN: urn:nbn:de:0183-mibe0002678

This article is freely available from
https://doi.org/10.3205/mibe000267

Published: *2024-08-09*