

**SePaMiM** - Sequential Pattern Mining und  
Pattern Matching von Krankheits- und  
Behandlungsverläufen für klinische Krebsregister

Gefördert vom G-BA Innovationsfond  
FKZ: 01VSF20018

# KONZEPTION EINES ANALYSEMODELLS UND ABFRAGEWERKZEUGS

27. JULI 2023

KOLJA BLOHM, DAVID KORFKAMP  
OFFIS – Institut für Informatik  
Escherweg 22, 26121 Oldenburg

|                                       |                                                                                                                 |
|---------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| <b>GESAMTPROJEKTL EITUNG</b>          | PROF. DR.-ING. ANDREAS HEIN<br>OFFIS – INSTITUT FÜR INFORMATIK                                                  |
| <b>PROJEKTL EITUNG DER EVALUATION</b> | FLORIAN OESTERLING<br>LANDESKREBSREGISTER NRW, BOCHUM                                                           |
| <b>KONSORTIALFÜHRUNG</b>              | OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG                                                                      |
| <b>KONSORTIALPARTNER</b>              | LANDESKREBSREGISTER NRW, BOCHUM                                                                                 |
| <b>KOOOPERATIONSPARTNER</b>           | OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE<br>/ EPIDEMIOLOGISCHES KREBSREGISTER<br>NIEDERSACHSEN, OLDENBURG |

## Inhalt

|   |                               |   |
|---|-------------------------------|---|
| 0 | Präambel .....                | 2 |
| 1 | Analysemodell.....            | 2 |
| 2 | Abfragewerkzeug .....         | 4 |
| 3 | Abfragesprache .....          | 4 |
| 4 | Ausführung von Anfragen ..... | 5 |

## 0 Präambel

Dieses Dokument beschreibt die Konzeption des Analysemodells und des Abfragewerkzeugs, die im Rahmen des Projekts SePaMiM erstellt wurde.

## 1 Analysemodell

In Zusammenarbeit mit dem LKR.NRW wurde ein Modell entworfen, das als Grundlage für die Suche nach Behandlungsverläufen dient. Dieses Analysemodell basiert auf den Best-Of-Datensätzen des LKR.NRW und der klinischen Landesauswertungsstelle für Krebsregisterdaten in Niedersachsen (KLast), die jeweils aus dem oBDS-Basisdatensatz abgeleitet wurden. Im Modell sind Patienten, Tumoren und Ereignisse abgebildet. Ein Patient hat einen oder mehrere Tumoren und jeder Tumor enthält ein oder mehrere Ereignisse. Ein Patient wird durch die Entität *Patient* dargestellt und enthält die Stammdaten des jeweiligen Patienten. Ein Tumor wird durch die Entität *Tumor* dargestellt und enthält Best-Of-Werte zum jeweiligen Tumor. Der wichtigste Aspekt in diesem Modell sind die Ereignisse, die den Behandlungsverlauf abbilden. Wir unterscheiden grundlegend zwei Typen von Ereignissen: Behandlungsereignisse, die die Ereignistypen *Operation*, *Strahlentherapie* und *Systemtherapie* umfassen, sowie Ereignistypen, die die Änderungen im Status des Patienten bzw. Tumors beschreiben. Hierunter fallen die Ereignistypen *Metastasen*, *Histologie*, *weitere Klassifikationen* und *Tumorstatus*.

Alle Ereignisse verfügen über einen Start- und einen Endzeitpunkt. Bei Ereignissen, die keine Dauer haben und somit Zeitpunktereignisse darstellen, wie z.B. die Meldung einer Metastase, sind Start- und Endzeitpunkt identisch. Bei Ereignissen, bei denen das Ende noch nicht bekannt ist, beispielsweise einer laufenden Bestrahlung, wird der Endzeitpunkt leergelassen.

Im Folgenden werden die Ereignistypen kurz erläutert:

*Operationen* stellen einen operativen Eingriff dar, der u.a. durch Listen von OPS-Codes und Komplikationen näher dargestellt wird. Das Ereignis *Strahlentherapie* fasst mehrere Bestrahlungen zusammen und stellt deren Intention dar. Es beinhaltet u.a. außerdem aufgetretene Nebenwirkungen. Das Ereignis *Systemtherapie* repräsentiert beispielsweise Chemotherapien, Immuntherapien oder Antihormontherapien. Wie bei der Strahlentherapie werden auch bei der Systemtherapie Nebenwirkungen dokumentiert.

Das *Metastasenereignis* dient der Dokumentation des Auftretens von Metastasen. Es enthält u.a. Angaben zur Lokalisation und zum Zeitpunkt der Diagnose der Metastase. Das *Histologieereignis* enthält z.B. Angaben zum Grading und zu den untersuchten Lymphknoten des Tumors. Neben der Abbildung klinisch relevanter Klassifikationen wie UICC oder TNM in den dafür vorgesehenen Feldern bietet der oBDS-Datensatz zusätzlich die Möglichkeit, Angaben in beliebigen, ansonsten nicht berücksichtigten Klassifikationssystemen zu melden. Solche Angaben werden im Datenmodell im Ereignis *weitere Klassifikationen* abgebildet. Der *Tumorstatus* dient dazu, den aktuellen Zustand des Tumors möglichst vollständig abzubilden. Dazu werden Informationen, die im oBDS-Datensatz in verschiedenen Entitäten, wie z.B. Operationen oder Verlaufsmeldungen angegeben werden können, zusammengefasst. Dazu gehören z.B. TNM-Angaben sowie Angaben zu Remission, Progression und Rezidiv. Die Angaben in einem Tumorstatus-Ereignis werden sukzessive fortgeschrieben, bis sie durch

Informationen aus neueren Meldungen aktualisiert werden. Beginn- und Enddatum des Tumorstatus geben den Zeitraum an, in dem die Informationen im Tumorstatusereignis gültig sind. Auf diese Weise wird ein Tumor von der Diagnose bis zum möglichen Tod des Patienten durchgehend durch die in den Tumorstatus-Ereignissen enthaltenen Informationen beschrieben. Die Abbildung des aktuellen Zustands des Tumors in einem hierfür dedizierten Ereignis dient der Vereinfachung von Auswertungen, da nicht mehr manuell nach aktuell gültigen Werten gesucht werden muss. Die Entitäten des Datenmodells sind in Abbildung 1 dargestellt. Bei Bedarf lässt sich das Analysemodell einfach erweitern.

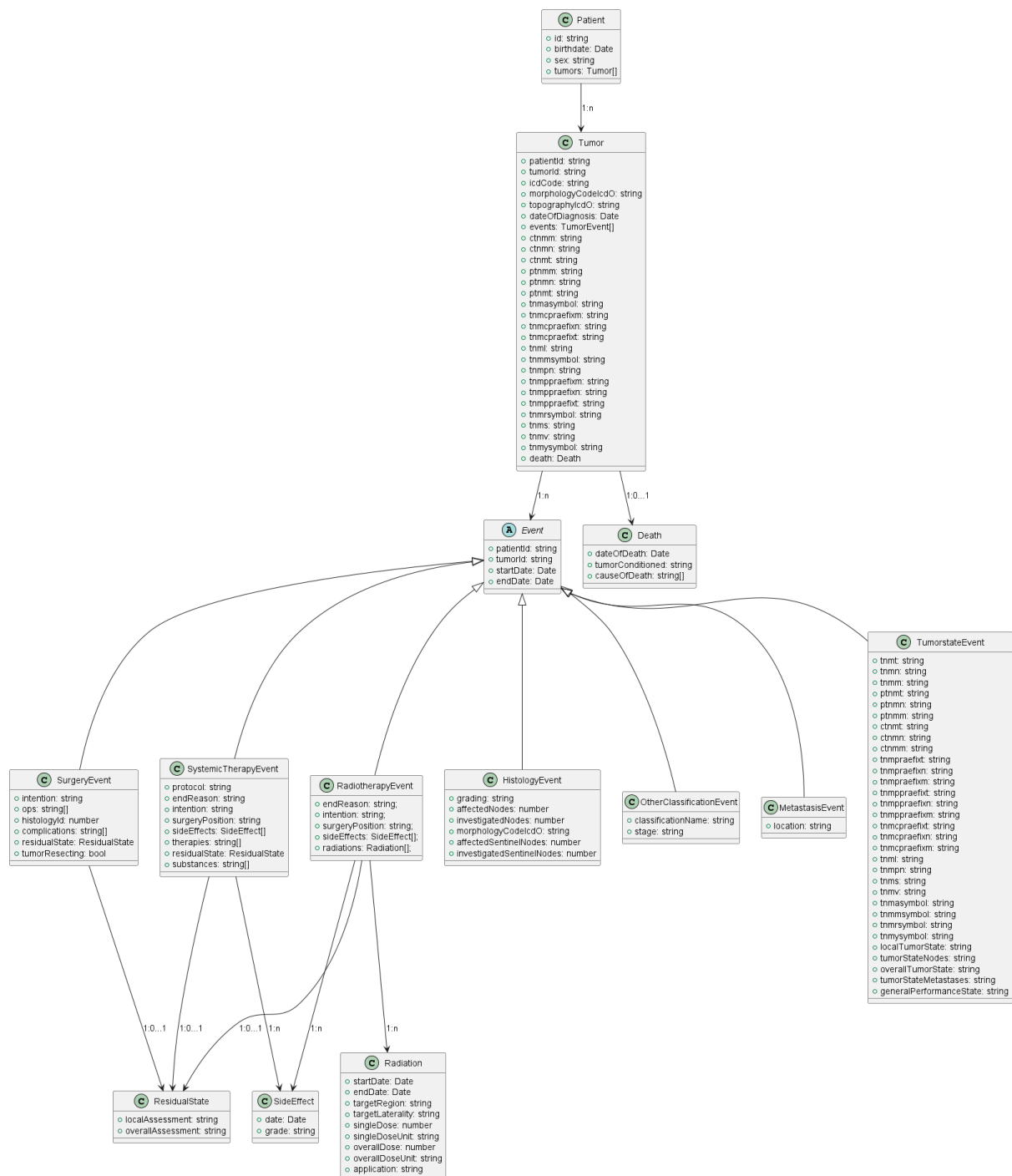


Abbildung 1 Entitäten des Analysemodells

## 2 Abfragewerkzeug

Aufbauend auf den in Meilenstein 4 beschriebenen Anforderungen wurde ein Konzept eines Abfragewerkzeugs erstellt. Bei dem Konzept handelt es sich um ein Grobkonzept, das im Laufe des Projektverlaufs voraussichtlich überarbeitet wird.

Das Abfragewerkzeug ist als eine Client-Server-Anwendung konzipiert. Anfragen sollen über einen Webclient, beispielsweise Chrome oder Firefox, gestellt werden. Der Webclient kommuniziert über eine REST-API mit einem Backend-Server, der die Anfragen verarbeitet. Die Ergebnisse der Anfragen werden dann im Webclient präsentiert.

Die Daten für das Abfragewerkzeug werden in einer Datenbank gespeichert. Der derzeitige Stand sieht die Nutzung einer Postgres Datenbank vor. In diesem wird das oben beschriebene Analysemodell abgebildet. Alle Ereignisse werden in einer Event-Tabelle gespeichert. Diese enthält den Start- und Endzeitpunkt, sowie die Art des Ereignisses und die ereignisspezifischen Daten. Die ereignisspezifischen Daten werden in einer JSON-Spalte als JSON-Dokument gespeichert, da sie sich je nach Ereignisart unterscheiden.

Einer der Gründe für die Entscheidung für Postgres ist es, dass es sich hierbei um eine Open Source Datenbank handelt, die frei verfügbar und seit geraumer Zeit etabliert ist. Durch Ihre Plattformunabhängigkeit lässt sie sich in den diversen Systemumgebungen der Krebsregisterlandschaft einfach in Betrieb nehmen. Zudem ist Postgres beim LKR.NRW bereits im Einsatz. Das Speichern der Ereignisse als JSON-Dokumente hat mehrere Vorteile:

- Alle Daten eines Ereignisses sind in einem Dokument gespeichert und müssen nicht durch das Joinen mehrerer Relationen zusammengeführt werden.
- Das Schema lässt sich einfach erweitern, auch ohne das Hinzufügen weiterer oder Anpassen bestehender Tabellen.
- Für die spätere Verarbeitung von Abfragen im Hauptspeicher werden vollständige Behandlungsverläufe benötigt, so dass Ereignisse in jedem Fall vollständig abgefragt werden müssen.

Nachteile dieses Ansatzes sind die bekannten Probleme, die bei der Denormalisierung von Daten auftreten, wie z.B. höherer Speicherplatzbedarf und die Möglichkeit von Dateninkonsistenzen. Da wir bei der Datenintegration das Analysemodell aus einer normalisierten Form generieren, lassen sich Dateninkonsistenzen einfach vermeiden. Der zusätzliche Speicherbedarf ist ein nachrangiges Problem, da die Datenmenge, die für die Evaluation und auch in einem späteren Produktiveinsatz zu erwarten ist, nur wenige GB umfasst.

## 3 Abfragesprache

Für die Formulierung von Abfragen an das Analysemodell wird die Clinical Quality Language (CQL) genutzt. CQL wird vom HL7-Konsortium entwickelt, und dient dazu, Patientenkohorten zu definieren und abzufragen. CQL bietet insbesondere verschiedene Konstrukte, um Abfragen auf Verlaufsdaten zu formulieren. Hierzu gehört beispielsweise die Unterstützung von Intervalldaten und geeigneten Operatoren, um Intervalle miteinander zu vergleichen. CQL ist eine Abfragesprache, die speziell für den medizinischen Kontext entwickelt wurde und

insbesondere im US-amerikanischen Raum bereits breiten Einsatz findet. Entsprechend scheint die Anwendung von CQL auf klinischen Krebsregister Daten vielversprechend.

## 4 Ausführung von Anfragen

Für das Ausführen von CQL-Abfragen wird die Open-Source CQL-Bibliothek<sup>1</sup> der Clinical Quality Framework Initiative genutzt. Über entsprechende Programmierschnittstellen ermöglicht die Bibliothek die Erweiterung um eigene Datenquellen und Datenmodelle. Das genutzte CQL-Datenmodell basiert auf dem oben beschriebenen Analysemodell. Die Postgres-Datenbank dient als Datenquelle. Beim Start des Servers werden die Daten aus der relationalen Datenbank in Java-Objekte eingelesen und für spätere Anfragen im Speicher vorgehalten. Neben Ausführung der CQL-Anfragen können die Ergebnisse im Server für die Darstellung im Webclient aufbereitet werden. Beispielsweise kann der Server bei Bedarf eine Überlebenszeitanalyse berechnen.

---

<sup>1</sup> [https://github.com/cqframework/clinical\\_quality\\_language](https://github.com/cqframework/clinical_quality_language)