Ergebnisbericht (gemäß Nr. 14.1 ANBest-IF)

Konsortialführung:	OFFIS e.V.
Förderkennzeichen:	01VSF20018
Akronym:	SePaMiM
Projekttitel:	Sequential Pattern Mining und Pattern Matching von Krankheits- und Behandlungsverläufen für klinische Krebsregister
Autorinnen und Autoren:	OFFIS & LKR.NRW
Förderzeitraum:	01.04.2021 - 31.03.2024
Ansprechperson:	Dr.-Ing. Christian Lüpkes, OFFIS Escherweg 2, 26121 Oldenburg

Das dieser Veröffentlichung zugrundliegende Projekt SePaMiM wurde mit Mitteln des Innovationsausschusses beim Gemeinsamen Bundesausschuss unter dem Förderkennzeichen 01VSF20018 gefördert.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Zusammenfassung

Hintergrund: Mit der Einrichtung der klinischen Landeskrebsregister und der Einführung eines bundesweit einheitlichen Datenschemas zur Meldung von einzelnen Krebsfällen, bezeichnet als onkologischer Basisdatensatz, werden Daten zur Behandlung und dem Erkrankungsverlauf von Krebspatient*innen in klinischen Krebsregistern erfasst. Das volle Potential dieser umfangreichen Verlaufsdaten wird bislang noch nicht erschlossen, da wenig Analysesoftware für komplexe Verlaufsdaten existiert. Derzeit sind Auswertungen nur mit hohem manuellen Aufwand umzusetzen. Zudem muss der zugrundeliegende, aus Einzelmeldungen zusammengefügte Datensatz kontinuierlich validiert und plausibilisiert werden. Bisher existieren dazu vor allem Kennzahlen, die die Diagnosedaten betrachten oder einzelne Behandlungsschritte, wie eine Operation. Für die Plausibilisierung des Behandlungsverlaufs im Ganzen fehlen geeignete Herangehensweisen.

Methodik: Im Rahmen des SePaMiM-Projekts wurden zwei Ansätze zur Analyse der Verlaufsdaten untersucht: Im Pattern Matching Ansatz wurden Möglichkeiten geprüft, auswertende Expert*innen in die Lage zu versetzen, zeitlich festgelegte Therapie- und Verlaufsereignisabfolgen als Suchkriterium vorzugeben, um Patient*innenkohorten zu selektieren. Im Data Mining Ansatz wurde durch die Verwendung von Clustering Verfahren die automatische Identifizierung von Kohorten anhand von Verläufen untersucht. Dazu wurden Entwicklungs- und Evaluationskonzepte erarbeitet. Die ingenieurswissenschaftlich implementierte Software (Pattern Matching Auswertewerkzeug PMAW und Data Mining Werkzeug DMW) wurden abschließend evaluiert und die Ergebnisse dokumentiert.

Ergebnisse: PMAW wurde anhand verschiedener Fragestellungen, die typische Aufgaben klinischer Krebsregister abbilden, evaluiert. Dabei dienten die bisher im Krebsregister verwendeten Vorgehen als Goldstandard und wurden mit den Ergebnissen des PMAW verglichen. Geringe Abweichungen ließen sich erklären. DMW wurde in drei Abschnitten evaluiert: 1. eine technische Evaluation, um zu zeigen, dass die Ergebnisse des Clusterings plausibel sind. 2. Vorstellung und Diskussion der Ergebnisse mit Fachexpert*innen des LKR.NRW (Das Clustering wurde als wertvoller Beitrag zur Unterstützung der Mitarbeiter*innen der klinischen Krebsregister bewertet). 3. Die durch das Clustering ermittelten Kohorten mit Überlebenszeitanalysen weiterberechnet und von den Fachexpert*innen inhaltlich als plausibel bewertet. Dafür standen Behandlungsverläufe von 1.058.706 Patient*innen mit insgesamt 7.688.816 Ereignissen zur Verfügung.

Diskussion: Die PMAW Evaluation zeigte, dass es verschiedene Routineauswertungen in der klinischen Krebsregistrierung unterstützt. Die Verwendung einer einheitlichen Definition und Datenmodells ermöglicht die Vergleichbarkeit von Auswertungen unterschiedlicher Krebsregister. Dies zeigte sich bspw. bei der Implementierung der Qualitätsindikatoren. Das Clustering mittels DMW konnte für die Fachexpert*innen viele interessante Einblicke in die Daten liefern. Beispielsweise konnten eine ganze Reihe von Clustern mit unerwarteten Verläufen identifiziert werden, die auf Datenqualitätsprobleme hinweisen können.

Schlagworte: Krebsregister, Data Science, Verlaufsdaten, Sequental Pattern Matching, Clustering

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Inhaltsverzeichnis

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

I Abkürzungsverzeichnis

IA	Innovationsausschuss Ab
G-BA	Gemeinsamer Bundesausschuss
§65c SGB V	Gesetz zur Einrichtung der klinischen Landeskrebsregister
ADT	Arbeitsgemeinschaft Deutscher Tumorzentren
ARMA	Autoregressive Moving Average
ATC	Anatomisch-Therapeutisch-Chemisches Klassifikationssystem
CEP	Complex Event Processing
CQL	Clinical Quality Language
CTC	Common Toxicity Criteria, systematische Einteilung von unerwünschten Ereignissen
DMW	In SePaMiM entwickeltes Data Mining Werkzeug zum Clustering
DTW	Dynamic Time Warping
EKN	Epidemiologisches Krebsregister Niedersachsen
ETL	Prozess der Extratktion, Transformation und Laden von Daten in ein Informationssystem
GEKID	Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V.
HL7	Health Level 7 ist eine Gruppe internationaler Standards für den Austausch von Gesundheitsdaten
ICD10	10. Version der Internationalen statistischen Klassifikation der Krankheiten und verwandter Gesundheitsprobleme
ICD-O-3	3. Revision der International Classification of Diseases for Oncology
KLast	Klinische Landesauswertungsstelle des Krebsregisters Niedersachsen
LKR.NRW	Landeskrebsregister NRW, Konsortialpartner
MDX	Multidimensional Expressions (MDX) sind eine komplexe Anfragesprache für multi-dimensional aufbereite Data Warehouses
NoSQL	Not only SQL; Datenbanken die einen nicht relationalen Ansatz verfolgen
oBDS	onkologischer Basisdatensatz, vorher als bundeseinheitlicher ADT-GEKID Basisdatensatz bezeichnet. Ein Schema zur Beschreibung der zu übermittelnden Datensätze.
OFFIS	Institut für Informatik - Oldenburg, Konsortialführer
OFFIS CARE	Betreiber der KLast sowie der Registerstelle (= Auswertung) des EKN, Kooperationspartner
OPS	Operationen- und Prozedurenschlüssel
PMAW	In SePaMiM entwickeltes Pattern Matching Werkzeug
QI	Qualitätsindikatoren
RPM	Row Pattern Matching
SePaMiM	Sequential Pattern Mining und Pattern Matching von Krankheits- und Behandlungsverläufen für klinische Krebsregister
SQL	Structured Query Language - Abfragesprache für relationale Datenbanken
TSH	Distance Threshold

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

UMAP	Uniform Manifold Approximation and Projection, Verfahren zur Dimensionsreduktion

II Abbildungsverzeichnis

III Tabellenverzeichnis

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

1 Projektziele

Mit der Einrichtung der klinischen Landeskrebsregister nach §65c SGB V und der Einführung eines bundesweit einheitlichen Datenschemas zur Meldung von Krebsfällen, bezeichnet als onkologischer Basisdatensatz (oBDS), werden neben den bisherigen Daten zu Inzidenz und Mortalität, die durch epidemiologische Krebsregister gesammelt und ausgewertet werden, nun auch die Daten zur Behandlung und dem Erkrankungsverlauf von Krebspatient*innen in klinischen Krebsregistern erfasst (1). Diese Anforderungen bedeuten für die klinischen Register, aber vor allem auch für die meldenden Einrichtungen (z.B. niedergelassene Onkologen, Pathologien oder Kliniken) eine erhebliche Anpassung der Erfassungsprozesse gegenüber den bereits etablierten Erfassungsprozessen der epidemiologischen Krebsregister. Neue Softwareschnittstellen und Meldeverfahren mussten etabliert werden und die deutlich gestiegene Mehrerfassung geschultert werden. Zudem ist die Komplexität der auszuwertenden Daten erheblich angestiegen. Aufgaben der Register sind, die Qualität der Daten zu sichern und auf ihrer Basis Aussagen zur Versorgungssituation von Krebspatient*innen zu treffen. Die gängigen Auswertungen, z.B. von Einzelparametern wie Tumorstadien oder im Rahmen vorgegebener und eng definierter Qualitätsindikatoren sind dabei wichtige Methoden, die heute schon vielfach eingesetzt werden. Das volle Potential der umfangreichen Datenlieferungen zu individuellen Krebsfällen kann damit aber nicht erschlossen werden. Zudem muss der aus vielen Einzelmeldungen bestehende und stetig wachsende zugrundeliegende Datensatz kontinuierlich validiert und plausibilisiert werden. Bisher existieren dazu vor allem Kennzahlen, die die Diagnosedaten betrachten (z.B. Anteil der histologisch verifizierten Diagnosen) oder einzelne Behandlungsschritte, wie eine Operation. Für die Betrachtung von Behandlungsverläufen im Ganzen fehlen geeignete Herangehensweisen. Für all diese Aufgaben ist es wichtig, Kohorten von Patient*innen anhand ihrer Therapiepfade zu identifizieren und für weitere Analysen selektieren zu können. Dies ist eine Herausforderung, die aufgrund der Komplexität der Daten ohne geeignete Verfahren und Werkzeuge derzeit nur mit großem, unverhältnismäßigem manuellem Aufwand umzusetzen ist.

Primäres Ziel des Projektes war es zu untersuchen, inwieweit die Identifikation und Selektion komplexer Behandlungs- und Krankheitsverläufe aus Krebsregisterdaten mithilfe geeigneter Data Mining und Analyseverfahren möglich ist, um die im vorherigen Abschnitt genannten Limitationen zu überwinden.

Im *Pattern Matching Ansatz* war das Forschungsziel zu untersuchen, wie Anfragen zur Definition von Patient*innen-Kohorten anhand komplexer Behandlungs- und Krankheitsverläufe durch Expert*innen gestellt werden können. Hierbei sollen insbesondere zeitliche Beziehungen zwischen unterschiedlichen Elementen der Verlaufsdaten als Filterkriterium genutzt werden können.

Im *Data Mining Ansatz* sollten mithilfe von KI und Data Mining Verfahren automatisch Muster in den Verlaufsdaten identifiziert werden. Im Anschluss können die gefundenen Patient*innenkohorten der weiteren statistisch-epidemiologischen Auswertung zugeführt werden.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Aus dem Primärziel wurden eine primäre und vier sekundäre Forschungsfragen abgeleitet. Die primäre Forschungsfrage lautete:

- *Wie lassen sich die im onkologischen Basisdatensatz enthaltenen Behandlungsverläufe aufbereiten und analysieren, so dass sie zur Unterstützung der Versorgungsforschung in klinischen Krebsregistern genutzt werden können?*

Die sekundären Forschungsfragen waren:

- *(2.1) Lassen sich aus den gewonnenen Erkenntnissen Standards für Routineauswertungen in den Krebsregistern etablieren?*
- *(2.2) Lassen sich alternative Behandlungs- und Krankheitsverläufe erkennen?*
- *(2.3) Wie lassen sich so gefundene Kohorten sinnvoll analysieren?*
- *(2.4) Lässt sich die Qualität der Daten bewerten?*

Im Rahmen der Projektdurchführung wurden zwei Evaluationskonzepte erstellt, die jeweils die Vorgehensweise und den Umfang der Evaluation von einem der beiden Ansätze beschreiben, um somit die Forschungsfragen beantworten zu können.

Zum Verständnis der Projektinhalte ist dabei wichtig, dass die Verfahren für Datensammlungen in Krebsregistern erarbeitet werden, die sich aus Meldungen im Format des onkologischen Basisdatensatzes (oBDS) zusammensetzen. Der oBDS und die Aufbereitung im Krebsregister sind in Abschnitt 2.3.2 beschrieben.

2 Projektdurchführung

2.1 Projektbeteiligte

Tabelle 1: Übersicht der Projektbeteiligten

Einrichtung	Verantwortlichkeit
Konsortialführung	
OFFIS e.V. - Institut für Informatik (OFFIS), Oldenburg Ansprechpartner: Dr.-Ing. C. Lüpkes (Gesamt-Projektleitung)	Steuerung und Management des Gesamtprojekts Untersuchung von geeigneten Daten Modellen und Auswertealgorithmen, Entwicklung der Daten Modelle und Adaption der Algorithmen in Form von zwei Softwarewerkzeugen und einer Datenbankstruktur.
Konsortialpartner	
Landeskrebsregister NRW gGmbH (LKR.NRW), Bochum Ansprechpartner: F. Oesterling (Verantwortliche Person für die Methodik)	Inhaltliche Krebsepidemiologie und fachliche Bewertung der IT-Technische Vorgehen und deren Evaluation. Bereitstellung von anonymisierten Daten
Kooperationspartner / Kooperativ Beteiligte	
OFFIS CARE GmbH (KLast), Oldenburg Ansprechpartner: Dr. Dr. J. Hübner	Fachliche Beratung zu Anwendungsgesichtspunkten und Auswertemöglichkeiten

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

2.2 Beschreibung/ Darstellung des Projekts

Mithilfe computergestützter Methoden und Verfahren des maschinellen Lernens sind die Möglichkeiten zur Auswertung von in Krebsregistern gespeicherten Daten in den vergangenen Jahren erheblich gestiegen; gleiches gilt aber auch für die Komplexität dieser Daten. Ihr volles Potenzial kann mit den gegenwärtigen Standards noch nicht vollständig erschlossen werden. Dazu wäre es wichtig, Behandlungs- und Krankheitsverläufe zu identifizieren und für weitere Analysen selektieren zu können. Ohne geeignete Verfahren und Werkzeuge ist diese Herausforderung nur mit enormem manuellem Aufwand umzusetzen. Bislang existieren vor allem Kennzahlen, welche die Diagnosedaten betrachten – so etwa den Anteil der histologisch verifizierten Diagnosen – oder einzelne Behandlungsschritte, wie beispielsweise eine Operation. Es fehlen geeignete Herangehensweisen, die einen Behandlungsverlauf im Ganzen plausibel verdeutlichen.

SePaMiM untersuchte deshalb, inwiefern die Identifikation und Selektion komplexer Behandlungs- und Krankheitsverläufe aus Registerdaten mithilfe geeigneter IT-Systeme unterstützt werden kann. Hierzu sollen geeignete Verfahren des Pattern Matching sowie Methoden des Data Mining genutzt werden. Dazu wurde zwei Demonstratoren (Software-Prototypen) einer IT-Anwendung entwickelt. Diese Demonstratoren wurden insbesondere an Datensätzen des Landeskrebsregisters Nordrhein-Westfalen erprobt, um Patient*innengruppen anhand der Behandlungs- und Krankheitsverläufe zu ermitteln und mit den Ergebnissen sowohl eine Qualitätssicherung der Daten als auch eine vergleichende Analyse der Gruppen durchzuführen.

2.3 Beschreibung Ablauf des Projekts

Das Projekt wurde auf eine Laufzeit von 3 Jahren geplant. Die inhaltliche Aufgabenteilung in diesem Projekt lässt sich folgendermaßen unterteilen: Das Landeskrebsregister NRW stellte die Fachexpertise bereit, mit der es die Anforderungen definierte, während der Implementierung Feedback gab und zusammen mit dem OFFIS die Evaluation durchführte. Das OFFIS war technischer Partner, der primär für die Planung, Entwicklung der Analysefunktionalitäten und Umsetzung der Demonstratoren zuständig war. Die OFFIS CARE, als Betreiber der Klinischen Auswertungsstelle des Krebsregisters Niedersachsen, unterstützte die Entwicklung erster Lösungsideen und Umsetzungen aus der Anwenderperspektive.

Die Vorgehensweise zur Umsetzung des Pattern-Matching-Ansatzes und des Data Mining Ansatzes war jeweils folgende: zuerst wurde im Rahmen einer Literaturrecherche der aktuelle Stand der Technik ermittelt. Hierauf aufbauend wurde ein Konzept zur Entwicklung eines Demonstrators erstellt. Außerdem wurde ein Konzept für die Evaluation des Demonstrators geschrieben. Anschließend wurde der Demonstrator implementiert und evaluiert.

Die Umsetzung der Demonstratoren erfolgte im Rahmen einer agilen Vorgehensweise bei der durch kontinuierliche Feedbackschleifen sichergestellt wurde, dass der Demonstrator die Anforderungen erfüllt.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

2.3.1 Phasenbasierter Arbeitsplan

Das Projekt war in sieben unterschiedliche Phasen unterteilt. In der ersten Phase wurden primär organisatorische Aufgaben durchgeführt. Dies umfasst insbesondere die Erstellung eines Datenschutz- und Übermittlungskonzepts um die Verfügbarkeit der Krebsregisterdaten des LKR.NRW für das Forschungsprojekt rechtlich abzusichern. In der zweiten Phase wurden inhaltliche Fragestellungen und Anforderungen an die zu erstellenden Systeme erhoben. In Phase 3 wurde mit der Planung und Vorbereitung des Pattern-Matching-Ansatzes begonnen. Hierbei wurden unter anderem eine Literatur- und Internetrecherche zur Ermittlung des derzeitigen Stands der Technik durchgeführt und die spätere Evaluation vorbereitet. In der darauffolgenden Phase 4 wurde das Pattern-Matching-Werkzeug (PMAW) implementiert und getestet. Parallel dazu wurde in den Phasen 5 und 6 das Konzept für das Data Mining erarbeitet und (als DMW) umgesetzt, mit analogen Arbeitsschritten zu denen aus Phasen 3 und 4. Die letzte Phase beinhaltete die Evaluation beider Ansätze.

Die konkrete Durchführung war demnach wie folgt. Zunächst wurden in Phase 1 die rechtlichen und organisatorischen Rahmenbedingungen geklärt, damit das LKR.NRW die zu analysierenden Daten in einer geeigneten, anonymisierten Form für das OFFIS bereitstellen kann. Dies war notwendig, damit OFFIS ein tiefergehendes Verständnis der Datenbeschaffenheit mit Blick auf potentielle Fragestellungen bilden kann. Zudem diente es auch dazu, Prototypen so zu entwickeln, dass sie ohne weitere Anpassung für eine Evaluation im LKR.NRW angewendet werden können.

Ausgehend von den potentiell verfügbaren Daten wurden inhaltliche Fragestellungen vom LKR.NRW definiert, wobei OFFIS die informationstechnische Machbarkeit einschätzte. Nachdem diese wichtige Grundlage geschaffen war, wurden Algorithmen ausgewählt und die Daten analysegerecht modelliert. Die beiden prototypischen Anwendungen wurden dann dem LKR.NRW und der KLast regelmäßig präsentiert und deren Rückmeldungen integriert.

Das LKR.NRW entwickelte parallel zwei Evaluationskonzepte um die beiden Prototypen am Schluss nach definierten Kriterien zu bewerten.

2.3.2 Beschreibung der genutzten IT-Standards

Mit der Einrichtung der klinischen Landeskrebsregister nach §65c SGB V wurde für die einheitliche onkologische Dokumentation von Krebsfällen der bundesweit einheitliche onkologische Basisdatensatz (oBDS, ehemals ADT-GEKID-Basisdatensatz) eingeführt (1,2). Trotz der Bezeichnung „Datensatz“, handelt es sich beim oBDS nicht um einen Datensatz, sondern um ein Datenschema. Der oBDS ist ein Dokumentationsstandard zur Meldung von Krebsfällen der von allen deutschen Krebsregistern verpflichtend genutzt wird. Technisch ist die Übertragung von oBDS-Meldungen in einem XML-Datenformat vorgesehen. Die Struktur des oBDS wird dabei durch ein XML-Schema spezifiziert. Eine oBDS-Meldung kann Informationen zur Tumordiagnose, zur Behandlung und zum Krankheitsverlauf enthalten. Hierunter fallen beispielsweise Angaben zu Patient*innen, Stammdaten, Operationen, Radiotherapien oder auch das Auftreten von Fernmetastasen. Der oBDS nutzt verschiedene Klassifikationssysteme wie z.B. die ICD10, ICD-O-3, OPS, CTC, ATC und weitere. Der oBDS wird kontinuierlich weiterentwickelt und erweitert. Die aktuelle Version des oBDS, in der einzelne

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Entitäten, Felder und valide Wertebereiche beschrieben werden, kann unter https://basisdatensatz.de/ eingesehen werden.

Die Landeskrebsregister erhalten im Laufe der Zeit mehrere oBDS-Meldungen von Kliniken und Praxen zu einem Krebsfall. Ausschlaggebend für die Übermittlung ist die Erfüllung des jeweiligen Meldeanlasses in der jeweiligen Einrichtung. Die einzelnen Meldungen sind auch nach der Zuordnung zu einer Person bzw. einem Tumor im Krebsregister noch nicht unmittelbar auswertbar. Um einen geeigneten Auswertungsdatensatz zu erzeugen, ist eine weitere Aufbereitung der Einzelmeldungen erforderlich. Dabei wird versucht, die besten Informationen zu einem Erkrankungsfall sowie den einzelnen klinischen Ereignissen zu bestimmen. Dieser Prozess wird auch als Best-of-Bildung bezeichnet, da hier versucht wird, die beste Information zu einem klinischen Ereignis zusammenzufassen. Der resultierende Auswertungsdatensatz wird auch als Best-of-Datensatz bezeichnet. Der Prozess zur Best-of-Bildung lässt sich in zwei Teilprozesse unterteilen:

1. Ereigniszuordnung: Anhand von Matching-Regeln wird bestimmt, ob mehrere Meldungen dasselbe klinische Ereignis beschreiben. Dieser Teilprozess bestimmt die Anzahl von Ereignissen, die aus einer Menge von Meldungen generiert werden.
2. Merkmals-Best-of: Bei Erfüllung der Matching-Regeln werden konkurrierende Informationen aus den verschiedenen Meldungen regelbasiert zusammengefasst.

Eine häufige Situation ist die Meldung desselben Ereignisses durch mehrere Melder. Für die Zusammenführung der Informationen aus mehreren Meldungen (Merkmals-Best-of) gibt es Regeln, die festlegen, wie mit unterschiedlichen oder widersprüchlichen Informationen umzugehen ist, um die höchstwertigen Informationen aus den Meldungen zu extrahieren. Dabei wird z.B. berücksichtigt, wie präzise, plausibel und aktuell die einzelnen Informationen sind. Die einzelnen Regeln für die Ereigniszuordnung und das Merkmals-Best-of werden in einer krebsregisterübergreifenden Arbeitsgruppe (AG Bildung klinisches Best-of) auf Ebene der Plattform § 65c erarbeitet. Ziel ist es, eine einheitliche Best-of-Bildung in allen Landeskrebsregistern zu erreichen.

Die in den einzelnen Landeskrebsregistern erstellten Best-of-Datensätze werden jährlich in einem standardisierten Datenformat (oBDS-RKI) dem ZfKD zur Verfügung gestellt und dort zu einem bundesweiten Datensatz zusammengeführt.

Im Rahmen des Projekts wurde auf Best-of-Daten des LKR.NRW, sowie der KLast gearbeitet, welche auf den gesammelten, zusammengeführten Meldungen im oBDS-Format basieren.

Der Pattern Matching Ansatz wurde unter Einbeziehung der Abfragesprache CQL umgesetzt. CQL ist ein Standard der HL7 zur Abfrage von klinischen Daten (3).

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

3 Methodik

Der Forschungsschwerpunkt des Projektes beinhaltete die Operationalisierung von Daten zu Behandlungs- und Erkrankungsverläufen aus Best-of-Daten klinischer Krebsregister mithilfe geeigneter Analyseverfahren für sequentielle Daten. Hier wurden zwei primäre Forschungsrichtungen identifiziert: Die erste Forschungsrichtung hat das Ziel, in den Verlaufsdaten nach Sequenzen zu suchen, die einem vorgegebenen Suchmuster entsprechen (Pattern Matching). Die zweite Forschungsrichtung kommt aus dem Data Mining und befasst sich mit Algorithmen zum Finden von interessanten oder häufig auftretenden Mustern in den Sequenzdaten (Data Mining). In diesem Abschnitt beschreiben wir das methodische Vorgehen zur Bearbeitung der beiden Forschungsfragen.

3.1 Methodik Pattern Matching

Im Folgenden wird das methodische Vorgehen zur Erstellung und Evaluation des Pattern Matching Werkzeugs (PMAW) vorgestellt.

Zu Beginn des Projekts wurde damit begonnen, den notwendigen Wissensstand für die weitere Forschung und Entwicklung aufzubauen. Zunächst wurde mit der Einarbeitung in die zur Verfügung stehenden Daten begonnen. Hierzu wurde der durch oBDS definierte Eingangsdatensatz sowie die hierauf basierenden, für Auswertungszwecke aufbereiteten, Best-of-Datensätze der KLast und des LKR.NRW näher betrachtet und miteinander verglichen.

Um einen Überblick über mögliche Anwendungsszenarien für das PMAW zu bestimmen, wurden parallel hierzu typische Fragestellungen aus der Krebsregistrierung erarbeitet.

Dazu gehörte zum einen die Frage nach spezifischen, differenzierten Behandlungsverläufe, wie sie teilweise im Rahmen von Datenanträgen oder Qualitätsindikatoren (QIs) erforderlich sind. Solche Fragestellungen beinhalten das Filtern nach patienten- und tumorspezifischen Kriterien sowie die Identifikation eines Startereignisses (z. B. eine Therapie oder das Auftreten der ersten Fernmetastase) und eines oder mehrerer darauffolgender Ereignisse. Bei der Erstellung eines QIs dienen diese Kriterien zur Bestimmung von Zähler und Nenner des Indikators. Im Rahmen eines Datenantrags werden in der Regel alle beantragten Daten zu den betroffenen Fällen in pseudonymisierter Form bereitgestellt.

Darüber hinaus gibt es zahlreiche Fragestellungen, die der Überprüfung und Verbesserung der Datenqualität im Krebsregister dienen, darunter die Identifikation von Implausibilitäten (z. B. die gleichzeitige Verabreichung bestimmter Medikamente und einer Strahlentherapie).

Anhand dieser Fragestellungen wurde dann ein initialer Anforderungskatalog für das PMAW erstellt, der den Rahmen des Demonstrators festlegte (siehe Anlage 4, Anforderungskatalog PMAW). Ausgehend auf der Einarbeitung in die Datengrundlage und dem Anforderungskatalog wurde ein initiales Datenmodell erstellt, welches die Struktur der Daten beschreibt, die für spätere Analysen genutzt werden sollen. Das Datenmodell beinhaltet zwei Hauptbestandteile: die Best-of-Daten eines Tumors zum Zeitpunkt der Diagnose sowie die Best-of-Daten zum Behandlungsverlauf. Für die Übernahme der Daten aus dem LKR.NRW und der KLast in das Datenmodell wurden entsprechende ETL-Skripte zur analysegerechten Datenbereitstellung entwickelt. Zu Test- und Entwicklungszwecken hat das LKR.NRW dem

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

OFFIS anonymisierte Echtdaten bereitgestellt. Hierdurch konnte die ursprünglich geplante Generierung von künstlichen Testdaten entfallen.

3.1.1 Stand der Technik

Um den aktuellen Stand der Technik zu geeigneten Technologien und Verfahren zur Anfrage auf Verlaufsdaten zu erfassen, wurde eine Literatur- und Internetrecherche durchgeführt.

Die Beschreibung orientiert sich an der von vom Brocke et al. (4) beschriebenen Vorgehensweise zur Literaturrecherche im Bereich der Informationssysteme. Das Vorgehen ist in mehrere Phasen unterteilt, darunter die Definition des Review Scopes, die Konzeptionalisierung des untersuchten Themas, die Literatursuche und die Analyse der gefundenen Ergebnisse. Für SePaMiM wurden zunächst relevante Technologiearten identifiziert und der Suchraum auf diese eingegrenzt. Bei der Suche wurde iterativ vorgegangen und die Ergebnisse regelmäßig mit dem LKR.NRW diskutiert, um die Suchstrategie und bei Bedarf den Suchraum entsprechend kontinuierlich anzupassen. Dies ermöglichte es, einen praxisorientierten Überblick zu erhalten, der direkt in die Entwicklung der Softwareprototypen einfließen konnte.

3.1.1.1 Definition des Review Scopes

Das SePaMiM-Projektteam hat die Recherche mit dem primären Ziel durchgeführt, sich einen Überblick über mögliche Technologien und Verfahren zu verschaffen, die zur Suche auf Patient*innendaten in klinischen Krebsregistern geeignet erscheinen. Der Fokus lag hierbei insbesondere auf der Suche anhand von Behandlungsverläufen. Um den aktuellen Stand der Technik zu geeigneten Technologien und Verfahren zur Anfrage auf Verlaufsdaten zu erfassen, wurde eine Literatur- und Internetrecherche durchgeführt. Dabei wurden verschiedene Technologiearten anhand exemplarischer Vertreter betrachtet. Die Ergebnisse der Recherche dienen als Grundlage für die Konzeption und Implementierung der Suchfunktionalität im PMAW.

3.1.1.2 Konzeptualisierung des Themas

Auf Grundlage der definierten Anforderungen an das PMAW wurden potentiell geeignete Technologiearten identifiziert. Hierunter fielen relationale Datenbanken, NoSQL Datenbanken und Systeme, die speziell für den Einsatz auf medizinischen Daten entworfen wurden. Die Anforderungen wurden in eigenen Dokumenten erfasst. Zusammenfassend sind darin folgende Anforderungen enthalten:

- Patient*innen- und Diagnosedaten sowie Verlaufsereignisse müssen anhand ihrer Attribute gefiltert werden. Die zu filternden Entitäten und Ereignistypen können verschiedene numerische, kategorische und boolesche Attribute sowie verschachtelte Objekte und Listen enthalten.
- Verlaufsdaten müssen anhand ihrer zeitlichen Beziehungen gefiltert werden können. Dies umfasst sowohl die Reihenfolge von Ereignissen als auch die zeitlichen Abstände zwischen Ereignissen (z.B. eine Radiotherapie muss innerhalb von 90 Tagen nach

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Operation erfolgen). Dieser Aspekt ist besonders wichtig, da das Ziel des PMAW insbesondere die Abfrage anhand von Behandlungsverläufen ist.

- Eine für die Formulierung von Anfragen genutzte Anfragesprache soll nach Möglichkeit auch von Anwender*innen mit wenig technischem Hintergrund verstanden oder sogar selbstständig genutzt werden können.

Im Bereich der relationalen Datenbanken wurden zwei Ansätze der Kohortendefintion betrachtet: die Definition von Kohorten mit Hilfe von klassischen SQL-Konstrukten (wie Selektion, Projektion und Joins), sowie die Verwendung von Row Pattern Matching (RPM).

Beim RPM handelt es sich um einen relativ neuen Ansatz, der jedoch in letzter Zeit immer mehr an Bedeutung gewonnen hat und 2016 in den SQL Standard integriert wurde (5). RPM ermöglicht es, nach bekannten Mustern in sequentiellen Daten zu suchen. Muster werden mit Hilfe von Musterabfragen definiert, die der Definition von regulären Ausdrücken ähneln. Diese Abfragen definieren Beschränkungen für einzelne Zeilen sowie eine zeitliche Reihenfolge zwischen den Zeilen.

Ähnliche Ansätze lassen sich in der Datenstromverarbeitung finden. Das Complex Event Processing (CEP) ermöglicht die Erkennung von Mustern in Datenströmen (6). RPM hingegen ist für die Batchverarbeitung auf zum Anfragezeitpunkt vollständig vorliegenden Datensätzen gedacht. Grundlegende Konzepte sind jedoch oftmals übertragbar und werden auch in RPM Systemen genutzt. Beispiele für CEP Systeme sind SASE (7), Cayuga (8) und DejaVu (9). Aufgrund der Ähnlichkeit zu RPM Systemen und dem Fokus auf Datenstromverarbeitung haben wir CEP-Systeme frühzeitig ausgeschlossen.

Im Bereich der NoSQL Datenbanken wurden insbesondere dokumentenorienterte Datenbanken und Graphendatenbanken näher betrachtet. Dokumentenorientierte Datenbanken ermöglichen die Speicherung von semi-strukturierten Daten und scheinen daher gut für die Speicherung der z.T. sehr unterschiedlich strukturierten Behandlungsereignisse geeignet. Außerdem wurde untersucht, inwiefern sich zeitliche Verlaufsdaten in Graphen-Datenbanken abbilden und abfragen lassen. Es finden sich vereinzelt Ansätze bei denen Graphen-Datenbanken zur Analyse von medizinischen Daten zum Einsatz kommen (10).

Neben den zuvor beschriebenen Systemen, die sich Domänen-unabhängig einsetzen lassen, wurden auch unterschiedliche Systeme, die speziell für den Einsatz auf medizinischen Daten entworfen wurden, betrachtet.

3.1.1.3 Internet- und Literaturrecherche

Die Recherche verfolgte das Ziel, die einzelnen Technologiearten näher zu untersuchen. Dabei wurde kein Anspruch auf Vollständigkeit erhoben. Vielmehr sollte eine Übersicht über ein breites Spektrum an Technologien geschaffen werden, wozu die Eigenschaften und Fähigkeiten der einzelnen Technologiearten anhand exemplarisch ausgewählter Systeme betrachtet wurden. Bei Bedarf wurden einzelne Systeme einer vertieften Analyse unterzogen.

Unsere initiale Suche wurde primär über Google durchgeführt, um eine breite Basis an Informationen zu generieren, da Technologien oftmals nicht in wissenschaftlichen Publikationen, sondern auf Webseiten, Blogs und anderen Online-Ressourcen zu finden sind.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Abhängig von der Relevanz und Qualität der gefundenen Informationen wurden zusätzlich spezialisierte Datenbanken wie dbdb.io, PubMed und ResearchGate konsultiert. Die Suchbegriffe wurden iterativ verfeinert, um spezifischere und relevantere Ergebnisse zu erzielen. Dieser iterative Ansatz ermöglichte es uns, flexibel auf neue Erkenntnisse zu reagieren und die Suchstrategie kontinuierlich anzupassen. Die Ergebnisse der Recherche wurden regelmäßig mit dem LKR.NRW diskutiert, was gegebenenfalls zu weiteren Anpassungen der Suche führte. Im Folgenden wird die Suche für die einzelnen Technologiearten näher beschrieben.

Für die Suche nach relationalen und NoSQL-Datenbanken wurde neben Google insbesondere die von der Datenbankgruppe der Carnegie-Mellon-Universität gepflegte Datenbank (dbdb.io) mit 741 Datenbanksystemen herangezogen. Bei der Suche nach einzelnen Systemen wurden solche, die auf dbdb.io als "discontinued" markiert sind sowie Systeme, bei denen der letzte Commit auf GitHub mehr als ein Jahr zurückliegt, grundsätzlich nicht weiter berücksichtigt.

Für die Bewertung relationaler Datenbanken zur Kohortendefinition mit klassischen SQL-Konstrukten wurden drei weit verbreitete Datenbanken als Repräsentanten ausgewählt: Postgres, Microsoft SQL Server und Oracle. Der Grund für dieses vereinfachte Vorgehen liegt darin, dass sich SQL-Datenbanken in der Regel stark am SQL-Standard orientieren und sich daher insbesondere bei klassischen Selektions-, Filter- und Join-Operationen sehr ähneln.

Für die Suche nach Dokumenten-Datenbanken wurde der Filter für das Datenmodell in dbdb.io auf „document / XML“ gesetzt. Zusätzlich wurde Google genutzt, um weitere relevante Datenbanken zu finden. Dabei wurden Suchbegriffe wie „document database“ und „document store“ verwendet. Nach einer kurzen Betrachtung der GitHub-Seite und der Dokumentation wurden für die exemplarische Betrachtung folgende drei Systeme ausgewählt: MongoDB, ArangoDB und OrientDB. MongoDB wurde dabei näher untersucht, da es eine sehr verbreitete Dokumenten-DB ist. ArangoDB und OrientDB wurden aufgrund der ausführlichen Dokumentation und aktiven GitHub-Repositories ausgewählt. Weitere Systeme hätten ebenfalls in Betracht gezogen werden können, jedoch lag der Fokus darauf, einen repräsentativen Überblick zu schaffen und die Auswahl auf Systeme zu beschränken, die eine breite Akzeptanz und aktive Community-Unterstützung aufweisen.

Die Suche nach Graphendatenbanken geschah ähnliches wie bei den Dokumentendatenbanken: Der Filter für das Datenmodell in dbdb.io wurde auf „graph“ gesetzt, und für eine Google-Suche wurde insbesondere nach „graph database“ gesucht. Nach einer kurzen Betrachtung der GitHub-Seite und der Dokumentation wurden für die exemplarische Betrachtung die folgenden vier Systeme ausgewählt: Neo4j (11), AgensGraph (12), ArangoDB (13) und JanusGraph (14). Neo4j wurde aufgrund seiner Bekanntheit und weiten Verbreitung näher untersucht. AgensGraph und JanusGraph wurden aufgrund ihrer ausführlichen Dokumentation und aktiven GitHub-Repositories ausgewählt. ArangoDB, die auch schon als Dokumenten-Datenbanken ausgewählt wurde, wurde auch hier aufgenommen, als eine Datenbank die als multi-model Datenbank sowohl Aspekte einer Dokumenten- als auch Graphendatenbank aufweist.

Da es auf dbdb.io keine Möglichkeit gibt, die Datenbanken nach RPM-Unterstützung zu filtern, wurde die Liste der 741 Datenbanken durch eine Google-Suche weiter eingeschränkt. Dazu wurde der Datenbankname in Kombination mit jeweils einem der Begriffe „Row Pattern

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Matching“, „Row Pattern Recognition“ und „MATCH_RECOGNIZE“ als Suchbegriff verwendet. Bei einigen Datenbanken mussten Suchbegriffe ausgeschlossen werden, um relevante Ergebnisse zu erhalten. Sah in den ersten 20 Suchergebnissen mindestens ein Ergebnis vielversprechend aus, wurde die Datenbank genauer untersucht, um festzustellen, ob RPM unterstützt wird. Neben den Datenbanken auf dbdb.io wurden die in Bader, Kopp & Falkenthal (15) gelisteten Zeitreihen-Datenbanken mit in die Suche einbezogen. Außerdem wurde noch eine ergänzende Google Suche nach den Begriffen „Row Pattern Matching“, „Row Pattern Recognition“ und „MATCH_RECOGNIZE“, ohne Einschränkung auf eine bestimmte Datenbank, durchgeführt. Bei dieser Suche wurden die folgenden 11 Datenbanken, bzw. Tools mit RPM-Unterstützung ermittelt: Oracle (16), Snowflake (17), Trino (18), Hive (19), MADLib (20), ClickHouse (21), Vertica (22), Teradata (23), Flink (24), Calcite (25) und Beam (26). Da der Kern des SePaMiM Projekts die Analyse von Verlaufsdaten darstellt und Row Pattern Matching explizit für diese Art von Analyse gedacht ist, haben wir alle 11 Systeme in dieser Kategorie genauer betrachtet.

Für die Recherche nach *Anfragesprachen und Werkzeugen für die Suche auf medizinischen Daten* wurden die wissenschaftlichen Datenbanken PubMed und Researchgate sowie die Suchmaschine Google verwendet. Dabei wurden Suchbegriffe wie „clinical query language“, „medical query language“, „course of treatment query language“, „treatment course query language“, „clinical pathways query language“, „clinical data search“ oder „medical data search“ genutzt. Die Suche in PubMed wurde auf Publikationen ab dem Jahr 2010 eingeschränkt. Um einen Überblick zu erhalten, wurden Titel und Abstracts bzw. Kurzbeschreibungen einer Auswahl der gefundenen Artikel betrachtet. Relevante Artikel wurden bei Bedarf im Volltext gelesen. Ziel war es, möglichst etablierte Ansätze zu identifizieren. Bei der Suche wurden die folgenden Anfragesprachen und Werkzeuge als vielversprechende Kandidaten bzgl. der oben genannten Kriterien identifiziert und näher betrachtet: Arden Syntax (27), Clinical Quality Language (CQL) (3), i2b2 (28), OHDSI (29) und die Archetype Query Language (AQL) (30).

3.1.1.4 Analyse und Bewertung

In der folgenden Analyse und Bewertung der verschiedenen Technologiearten und Systeme werden die jeweiligen Vor- und Nachteile der betrachteten Technologien kurz zusammengefasst dargestellt. Es hat sich gezeigt, dass keine Technologie in allen Belangen den anderen überlegen ist.

Klassische SQL-Datenbanken zeichnen sich durch einen umfangreichen Funktionsumfang aus. Sie bieten vielfältige Möglichkeiten zur Filterung und Selektion von Daten. Außerdem können Daten mithilfe verschiedener Aggregations- und Gruppierungsfunktionen analysiert werden. Diese Eigenschaften machen sie zu einem mächtigen Werkzeug für die Datenverwaltung und -analyse. Allerdings sind SQL-Datenbanken für nicht-technische Anwender oft schwer zu nutzen, was auch die Erfahrungen des Krebsregisters NRW bestätigen. Die Komplexität der Abfragesprache und die Notwendigkeit, detaillierte Kenntnisse über die Datenbankstruktur zu haben, stellen für viele Anwender eine Hürde dar.

Dokumentendatenbanken eignen sich besonders für die Speicherung und Verwaltung komplexer, hierarchischer Datenstrukturen. Im Gegensatz zu relationalen Datenbanken, in

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

denen die Daten in der Regel über mehrere Tabellen verteilt werden, haben sie den Vorteil, dass sie oftmals keine Joins benötigen, um Daten einer Entität zusammenzuführen. Diese Eigenschaft kann dazu führen, dass verschiedene Abfragemuster, die in herkömmlichen relationalen Datenbanken nur schwer zu handhaben wären, vereinfacht werden. Da die Unterstützung für Joins in Dokumentendatenbanken eingeschränkt ist, ist die Verknüpfung von Daten, wie sie z.B. für das Filtern anhand von zeitlichen Beziehungen zwischen Ereignissen nötig ist, oftmals problematisch. Dies bedeutet, dass es beispielsweise nicht die Stärke solcher Datenbanken ist, zwei Ereignisse wie Operationen und Bestrahlungen miteinander in Beziehung zu setzen und anhand ihres Datums zu vergleichen. Solche Abfragen erfordern oft komplexe und ineffiziente Workarounds, was die Analyse und Auswertung der Daten erschweren kann.

Graphendatenbanken modellieren ihre Daten als Knoten und Kanten und sind daher grundsätzlich gut geeignet um Verlaufsdaten abzubilden. Die Formulierung von Anfragen wurde jedoch für nicht-technische Anwender*innen als schwierig und komplex empfunden. Während ein grundlegendes Verständnis der Datenstruktur auch in relationalen Datenbanken erforderlich ist, stellt die Modellierung in Graphendatenbanken eine besondere Herausforderung dar. Im Gegensatz zu den vertrauten Tabellenstrukturen relationaler Datenbanken sind die Konzepte von Knoten und Kanten für viele Anwender ungewohnt und erfordern eine Umstellung im Denkansatz. Zusätzlich erfordert die Nutzung von Abfragesprachen wie Cypher oder Gremlin spezialisierte Kenntnisse und setzen ein tiefgehendes Verständnis der spezifischen Syntax und Logik voraus, was die Benutzerfreundlichkeit weiter einschränkt.

Row Pattern Matching (RPM) ermöglicht es, nach vorgegebenen Mustern in sequentiellen Daten zu suchen. Muster werden mithilfe von Musterabfragen definiert, die der Definition von regulären Ausdrücken ähneln. Diese Abfragen definieren Filter für einzelne Zeilen sowie deren Reihenfolge. Einfache zeitliche Beziehungen, wie beispielsweise „eine brusterhaltende Operation soll vor einer Mastektomie auftreten“, lassen sich mit diesen Ansätzen sehr einfach abbilden. Komplexere zeitliche Beziehungen, wie beispielsweise zeitlich überlappende Therapien oder konkrete Zeitspannen zwischen zwei Ereignissen, werden jedoch meist nur schlecht oder gar nicht unterstützt. Insbesondere die Behandlung von Intervalldaten stellt eine Herausforderung dar. Da RPM hauptsächlich in relationalen Datenbanken umgesetzt wird, gelten für diesen Ansatz ähnliche Vor- und Nachteile wie für die o.g. klassischen SQL-Datenbanken. Dies bedeutet, dass RPM zwar mächtige Abfragemöglichkeiten bietet, aber für nicht-technische Anwender oft schwer zu nutzen ist. Zudem ist es bei der Einschränkung auf Reihenfolgen zwischen Zeilen oft schwierig, komplexere Datenstrukturen mit verschachtelten Daten zu filtern.

Bei den betrachteten Systemen, die dediziert für Anfragen von medizinischen Daten entwickelt wurden, fielen OHDSI Atlas und ib2b dadurch positiv auf, dass Anfragen primär in einer graphischen Oberfläche formuliert werden können. Dadurch lassen sich einfache Suchanfragen sehr einfach auch durch nicht-technische Anwender umsetzen. Für komplexere Anfragen, die beispielsweise Beziehungen zwischen Ereignissen abbilden, wurden diese Systeme jedoch als ungeeignet eingestuft.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Bei AQL, CQL und Arden handelt es sich um textuelle Anfragesprachen. Hierbei fiel CQL besonders positiv auf:

CQL ist eine vom HL7-Konsortium entwickelte Anfragesprache, die dazu dient, Patient*innenkohorten 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. CQL ist einfach zu verstehen und sehr ausdrucksstark. Außerdem unterstützt es neben unterschiedlichen Standard-Datenmodellen, wie FHIR auch die Definition eigener Datenmodelle, so dass hiermit ein Datenmodell basierend auf dem Analysemodell genutzt werden kann. Bei der Modellierung werden außerdem bei Bedarf Konzepte wie beispielsweise medizinische Codes und Maßeinheiten wie beispielsweise Milliliter und Zentimeter. Da Beziehungen zwischen Entitäten bereits im Datenmodell hinterlegt werden, können technische Konstrukte wie Joins in CQL Anfragen weitestgehend vermieden werden. Dies trägt zur Verständlichkeit von CQL bei.

Zwei Funktionen die CQL nicht unterstützt sind die Bildung von Aggregaten und das Stratifizieren von gefundenen Kohorten. Dies muss nachgelagert in einer statistischen Auswertungssoftware geschehen.

Zusammenfassend lässt sich feststellen, dass alle Systeme ihre Stärken und Schwächen zeigten und auch andere Umsetzungsmöglichkeiten denkbar und realistisch waren. So hätte man beispielsweise auch eine eigene Anfragesprache für Nutzende entwickeln und technisch in komplexe SQL-Anfragen übersetzen können. Da es sich bei CQL aber um einen in den USA bereits verbreiteten und etablierten Standard handelt, wurde trotz der Kenntnis der Limitationen entschieden, CQL für die Umsetzung des PMAW einzusetzen und zu prüfen, ob sich dieser Standard auch in Deutschland sinnvoll einsetzen lässt.

Entsprechend schien die Anwendung von CQL auf klinischen Krebsregister Daten vielversprechend. CQL wurde deshalb als Modellierungsbasis für das PMAW ausgewählt.

3.1.2 Konzeption

Basierend auf dem erstellten Anforderungskatalog und der Recherche zum Stand der Technik, wurde ein initiales Konzept zur Implementierung des PMAW entworfen. Das Konzept wurde im Rahmen der Entwicklung des PMAW regelmäßig anhand von Feedback aus Diskussionsrunden zwischen den Projektpartnern überarbeitet und angepasst.

Das PMAW verwendet CQL als Abfragesprache zur Definition von Kohorten. Die Struktur der für CQL-Abfragen verfügbaren Daten wurden in einem Datenmodell festgelegt. Das CQL-Datenmodell basiert auf dem oBDS-Datensatz sowie den Best-of-Datensätzen des LKR.NRW und der klinischen Landesauswertungsstelle Niedersachsen (KLast). Das Datenmodell kann über eine hierfür vorgesehene Software-Schnittstelle mit den Best-of-Daten verschiedener Krebsregister befüllt werden. Die Schnittstelle wurde für die Datenbestände des LKR.NRW und der KLast implementiert.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Das PMAW wurde als eine Client-Server-Anwendung konzipiert. Abfragen an das Datenmodell werden über ein Webinterface gestellt. Die Abfrageergebnisse werden visualisiert und können zusätzlich für weitere Verarbeitungsschritte exportiert werden.

Der Anforderungskatalog wurde nach Projektende um eine Bewertung des Erfüllungsgrads erweitert. Dieser wurde zu den einzelnen Anforderungen notiert. Alle Muss-Anforderungen wurden vollständig umgesetzt, einzelne Kann- und Soll-Anforderungen wurden nur teilweise bzw. wenige gar nicht umgesetzt.

3.1.3 Evaluation

Im Rahmen der Evaluation des PMAW wurde untersucht, ob sich CQL als Abfragesprache für die Beantwortung von typischen Fragestellungen aus der klinischen Krebsregistrierung eignet. Hierbei wurde außerdem geprüft, ob das CQL-Datenmodell alle für die Abfragen relevanten Informationen enthält. Das genaue Vorgehen wurde in einem Evaluationskonzept festgelegt.

Das Evaluationskonzept definierte initial fünf klinisch relevante Fragestellungen aus den Bereichen *Qualitätsindikatoren für Leitlinien und Plausibilitätsprüfung*, die mit Hilfe des PMAW beantwortet werden sollten. Das Konzept wurde im Verlauf des Projekts um vier weitere Fragestellungen aus den Bereichen *Externe Datennutzung und Routineauswertung* ergänzt. Die insgesamt neun unterschiedlichen Fragestellungen aus dem LKR.NRW und der KLast wurden mit CQL implementiert und die Ergebnisse mit den Ergebnissen aus bereits in den Krebsregistern etablierten Methoden verglichen. Hierbei wurden die gefundenen Tumore der beiden Anfragen miteinander verglichen und Precision und Recall berechnet. Bei Abweichungen wurden die Kohorten manuell abgeglichen und nach Unterschieden auf Einzelfallebene gesucht.

Im Folgenden werden die untersuchten Fragestellungen kurz vorgestellt.

Qualitätsindikatoren für Leitlinien sind Kennzahlen um die leitliniengerechte Behandlung von Krebspatient*innen zu bewerten. Die Berechnung von Qualitätsindikatoren anhand von Krebsregisterdaten wird von der Arbeitsgruppe QI der Plattform § 65c definiert. Es wurden exemplarisch die beiden Qualitätsindikatoren QI8 des Mammakarzinoms und QI6 des Lungenkarzinoms implementiert. QI8 beurteilt, den Anteil an Bestrahlungen nach einer brusterhaltenden Therapie beim Mammakarzinom. Qualitätsziel ist eine adäquate Rate an Bestrahlungen nach BET bei Patienten mit Ersterkrankung invasives Mammakarzinom. Der Anteil der Patientinnen, die eine adjuvante Strahlentherapie nach brusterhaltender Operation erhalten haben ist klinisch besonders relevant. Mehrere Metaanalysen haben belegt, dass die postoperative Bestrahlung effektiv lokoregionalen sowie entfernten Rezidiven vorbeugt (31,32). Es gibt keine Subgruppe von Patientinnen, die nicht von einer postoperativen Bestrahlung profitiert. Der zweite gewählte Qualitätsindikator, QI6 aus der S3 Leitlinie für das Lungenkarzinom beurteilt den Anteil an der Cisplatin-haltigen Chemotherapien nach einer OP bei Lungenkrebs. Hier ist das Qualitätsziel ist eine möglichst häufige Behandlung mit adjuvanter Cisplatin-basierter Chemotherapie bei NSCLC Stadium II b. Cisplatin ist das wichtigste Chemotherapeutikum in der Therapie des Lungenkarzinoms. Die Therapie mit Cisplatin ist laut einer Metaanalyse effektiver als die mit dessen Derivat Carboplatin und führt häufiger zu einer Remission beim metastasierten Lungenkarzinom (33). Daraus ergibt sich eine

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

unmittelbare klinische Relevanz für die Erhebung des Anteils an Cisplatin-basierten Chemotherapien für diese Krebsentität.

Im Bereich *Plausibilitätsprüfung* wurden drei Fragestellungen definiert: die Suche nach klinischen Ereignissen nach dem Tod eines Patienten, die Suche nach Patient*innen die im Laufe Ihrer Behandlung von ein einer palliativen zu einer kurativen Therapieintention gewechselt sind und die Suche nach Patient*innen bei denen ein Rückgang der Tumorgröße gemessen wurde, ohne das zwischen beiden Größenmessungen eine Therapie erfolgt ist. Da eine aussagekräftige Analyse der klinischen Fragestellungen nur gelingen kann, wenn die zugrundeliegenden Daten eine hohe Qualität aufweisen stehen bei diesen Fallbeispielen keine klinische Fragestellungen sondern die Prüfung der Datenqualität im Vordergrund.

Auf Anfrage können klinische Krebsregister Daten für Forschungszwecke an externe Forscher*innen bereitstellen. Im Bereich *Externe Datennutzung* wurden beispielhaft zwei Anträge für anonyme Einzelfalldaten, die an die KLast gestellt wurden, mit CQL umgesetzt. Hierbei wurden Daten für zwei Studien von einer Universitätsmedizin angefragt, um den Einfluss des Therapiebeginns bei fortgeschrittenem (UICC IV) kolorektalem, bzw. Mammakarzinom auf das Überleben zu untersuchen.

Krebsregister veröffentlichen regelmäßig aggregierte Krebsdaten. Diese Routineauswertungen dienen der Berichtspflicht gegenüber der Öffentlichkeit, der Politik und Gesundheitsplanung. Im Bereich *Routineauswertung* wurden exemplarisch zwei Seiten des interaktiven Onlineberichts der KLast mit CQL abgefragt. Die erste Seite enthält Daten über die Inzidenz verschiedener Diagnosen, stratifiziert nach Diagnosejahr, Geschlecht und Wohnort. Die zweite Seite stellt einen Überblick über durchgeführte Operationen dar. Diese werden nach unterschiedlichen Kriterien stratifiziert, u.a. Diagnosegruppe, Geschlecht, Wohnort und UICC-Stadium.

Die Ergebnisse der Evaluation werden in Abschnitt 4 "Projektergebnisse" beschrieben.

3.2 Methodik Data Mining

In diesem Abschnitt wird das methodische Vorgehen zur Erstellung und Evaluation des Data Mining Werkzeugs DMW zur Analyse von klinischen Behandlungsverläufen vorgestellt.

Ziel des *Data Mining Ansatzes war es,* mithilfe von Data Mining und Machine Learning Verfahren automatisch interessante Muster anhand der Behandlungsverlaufsdaten zu identifizieren, die dann im Anschluss von Expert*innen weiter analysiert und ausgewertet werden können. Im Folgenden wird das Vorgehen beschrieben, mit dem sich das SePaMiM-Team einen Überblick über potentiell geeignete Data Mining und Machine Learning Ansätze zur Analyse von Verlaufsdaten verschafft hat. Da das Ziel des Data Mining Ansatzes im Gegensatz zum PMAW-Ansatz bewusst sehr offen gehalten wurde, wurden vorab keine konkreten Anforderungen an das zu erstellende System gestellt. Daher wurden bei der Sichtung und Bewertung des Stands der Technik Ergebnisse und Ideen regelmäßige diskutiert und die Suche bei Bedarf in verschiedene Bereiche vertieft.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

3.2.1 Stand der Technik

Zunächst wurde recherchiert, welche Data Mining bzw. Machine Learning Ansätze zur Analyse von sequentiellen Daten existieren. Hierzu wurden beispielsweise Google und Researchgate nach Suchbegriffe wie „sequential data mining“, „sequence data mining“, „temporal data mining“ durchsucht. Dabei wurde kein Anspruch auf Vollständigkeit erhoben. Literatur- und Internetrecherchen haben dabei gezeigt, dass verschiedene Ansätze in Frage kommen. Bei den Ansätzen handelt es sich teils um Erweiterungen klassischer Data Mining Verfahren aber auch um Neuentwicklungen (35). Hierunter fallen z.B. folgende Ansätze:

- Klassifikation: Die Zuordnung von Sequenzen zu vordefinierten Klassen mittels überwachtem Lernen (35).
- Clustering: Die Gruppierung von Sequenzen basierend auf ihrer Ähnlichkeit, ohne vorherige Kenntnis der Gruppen (34).
- Sequential Pattern Mining: Entdeckung häufig auftretender Teilsequenzen in einer Menge von Sequenzen (36).

Die Ergebnisse der Recherche wurden regelmäßig gemeinsam zwischen OFFIS und LKR.NRW diskutiert. Da es sich bei den Daten der klinischen Krebsregister um einen neuartigen und sehr komplexen Datensatz handelt, ist das Auffinden neuer Strukturen besonders relevant. Daher wurde beschlossen, den Fokus auf unüberwachte Lernverfahren zu legen, da diese das Entdecken neuer Muster in den Daten ermöglichen. Von besonderem Interesse war dabei die automatische Erkennung von Patient*innen-Kohorten anhand ihrer Behandlungsverläufe. Hierzu erschien Clustering als besonders geeignet. In Folge dessen wurden noch einmal vertiefend die Eigenschaften von Clustering auf Verlaufsdaten untersucht und entschieden, den Fokus auf sequentielles Clustering als Data Mining Verfahren zu setzen.

3.2.2 Sequentielles Clustering

Das sequentielle Clustering befasst sich damit, eine Menge von Sequenzen anhand ihrer Ähnlichkeit in Gruppen einzuteilen (34). Im Rahmen des SePaMiM Projektes soll sequentielles Clustering dazu genutzt werden, Patient*innenkohorten anhand von Behandlungsverläufe zu bilden. Verfahren für das Clustern von Sequenzen lassen sich in drei grundlegende Kategorien einteilen (37,38): ähnlichkeitsbasiertes sequentielles Clustering, indirektes sequentielles Clustering und statistisches sequentielles Clustering. Diese werden in den folgenden drei Abschnitten kurz vorgestellt. Da die Behandlungsverlaufsdaten im SePaMiM Projekt aus komplexen, multidimensionalen Ereignissen bestehen, werden im Anschluss Möglichkeiten zum Clustern solcher Daten beschrieben.

3.2.2.1 Ähnlichkeitsbasiertes sequentielles Clustering

Ähnlichkeitsbasiertes sequentielles Clustering nutzt bereits vorhandene distanzbasierte Clustering Verfahren, wie zum Beispiel hierarchische Verfahren, dichtebasierte Verfahren, wie DBScan und partitionierende Verfahren, wie K-Means (37,38). Die Herausforderung besteht hierbei darin, ein geeignetes Ähnlichkeitsmaß zu finden, das die Ähnlichkeit zweier Sequenzen bestimmt.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Ein einfacher Ansatz, die Ähnlichkeit zweier Sequenzen gleicher Länge zu bestimmen, besteht darin, gegenüberstehende Elemente der Sequenzen zu vergleichen und die einzelnen Ergebnisse zu aggregieren.

Sequenz-Alignment Ansätze, sind dagegen nicht auf Sequenzen gleicher Länge beschränkt. Diese versuchen, Sequenzen optimal aufeinander abzubilden. Hierdurch berücksichtigen sie außerdem zeitliche Verschiebungen innerhalb der Sequenzen. Hierunter fallen z.B. Editierdistanzen und Dynamic Time Warping (DTW).
Editierdistanzen werden insbesondere für Sequenzen bestehend aus Symbolen genutzt. Sie bestimmen die Ähnlichkeit zwischen zwei Sequenzen anhand der minimal benötigten Modifikationen (Einfügen, Löschen und Ersetzen), um eine Sequenz in die andere zu überführen (37). Die Kosten für einzelne Modifikationen können dabei gewichtet werden, um zusätzlich Domänenwissen zu berücksichtigen.

DTW sucht die optimale Abbildung von zwei Sequenzen aufeinander. Hierbei werden jedoch keine Modifikationen vorgenommen, um eine Sequenz auf eine andere abzubilden. Stattdessen wird jedes Element einer Sequenz auf ein Element der Vergleichssequenz abgebildet. Es wird die Abbildung gesucht, die minimale Kosten verursacht, d.h. es werden möglichst ähnliche Elemente aufeinander abgebildet. Die Ähnlichkeit zwischen zwei Elementen kann dabei je nach Anwendungsfall unterschiedlich bestimmt werden. Bei der Abbildung wird außerdem eine Reihe von Einschränkungen berücksichtigt. Zum Beispiel kann nicht zurückgesprungen werden. Sobald ein Element der Eingangssequenz auf Element *n* der zu vergleichenden Sequenz abgebildet wurde, kann ein folgendes Element nicht mehr auf ein Element *n-1* oder kleiner der Vergleichssequenz abgebildet werden (39).

3.2.2.2 Indirektes sequentielles Clustering

Indirekte sequentielle Clustering Methoden verfolgen den Ansatz, Sequenzen zunächst in Feature-Vektoren umzuwandeln, um diese im Anschluss mit vorhandenen Clustering-Verfahren clustern zu können (37,38). Entsprechend wichtig ist hierbei die Auswahl geeigneter Features. Ein möglicher Ansatz besteht zum Beispiel darin, Sequenzen auf das Vorhandensein bzw. Nichtvorhandensein einzelner Elemente zu reduzieren. Alternativ kann auch die Häufigkeit der Elemente genutzt werden. Bei diesen Ansätzen geht jedoch die Reihenfolge der Elemente verloren. Eine Möglichkeit, die Reihenfolge der Elemente zu berücksichtigen, ist die Verwendung von kurzen Teilsequenzen. Eine Teilsequenz besteht aus *k* aufeinanderfolgenden Elementen und wird als Feature genutzt, das angibt ob bzw. wie oft eine Teilsequenz vorhanden ist (35). Bei numerischen Sequenzen können beispielsweise bekannte Verfahren aus der Signalverarbeitung, wie die Haar Wavelet Transformation genutzt werden, um die Sequenz in einen Feature-Vektor abzubilden (38,40).

3.2.2.3 Statistisches sequentielles Clustering

Beim statistischen sequentiellen Clustering werden statistische Modelle für das Clustering genutzt. Hierbei wird die Annahme getroffen, dass jede Sequenz durch ein statistisches Modell bzw. durch Wahrscheinlichkeitsverteilungen erstellt wurde (38). Die verschiedenen Ansätze lassen sich weiter anhand der Verwendung der statistischen Modelle in zwei Kategorien einteilen.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

In der ersten Kategorie wird davon ausgegangen, dass die den Sequenzen zugrundeliegenden Modelle jeweils ein Cluster darstellen. Entsprechend werden hierbei die Parameter der Modelle erlernt. Beispielsweise mit Hilfe des EM-Algorithmus (Expectation–maximization algorithm) (37). Die zweite Kategorie beinhaltet Methoden, die statistische Modelle zur Berechnung der Ähnlichkeit zwischen Sequenzen nutzen (38). Hierbei wird zunächst jede Sequenz auf ein Modell abgebildet und im Anschluss der Abstand zwischen zwei Sequenzen mittels Ähnlichkeitsmaßen für Wahrscheinlichkeitsverteilungen, wie z.B. der Kullback-Leibler-Divergenz ermittelt (38). Zum Clustern werden dann auf Abstandsmaßen basierende Verfahren genutzt. In beiden Ansätzen werden u.A. Markov Ketten, Hidden Markov Modelle oder Autoregressive Moving Average (ARMA) Modelle als statistische Modelle genutzt (37,38).

3.2.2.4 Umgang mit komplexen sequentiellen Daten

Viele Ansätze zum Clustern von sequentiellen Daten sind zunächst darauf ausgelegt, auf einfachen Sequenzen bestehend aus numerischen oder kategorialen Werten angewendet zu werden. Die Behandlungsverläufe, die im SePaMiM Projekt als Eingabe für das Clustering dienen, bestehen jedoch aus komplexen, multidimensionalen Ereignissen, die sowohl numerische als auch kategoriale Attribute besitzen. Zudem unterscheidet sich die Struktur der Ereignisse je Ereignistyp. Beispielsweise beinhalten Operationsereignisse u.a. eine Intention, eine Liste von Komplikationen, sowie eine Liste von OPS Codes. Strahlentherapien hingegen beinhalten u.a. ebenfalls eine Intention und zusätzlich je eine Liste von Bestrahlungen und von Nebenwirkungen.

Eine Möglichkeit, mit Sequenzen komplexer Daten umzugehen, besteht darin, die einzelnen Elemente zu kategorisieren und auf einfache Symbole abzubilden. Dieses Vorgehen wird zum Beispiel in Vogt et al. beschrieben (41). Die Autor*innen untersuchen die Versorgungspfade unterschiedlicher Patient*innen mit Herzinsuffizienz. Datengrundlage sind hierbei Abrechnungsdaten von Krankenkassen aus einer Periode von 24 Monaten. Aus diesen Daten werden einfache, aus Symbolen bestehende Sequenzen gebildet. Da die Abrechnungsdaten lediglich quartalsweise erhoben wurden, repräsentiert jedes Symbol einer Sequenz ein Abrechnungsquartal. Für jeden der drei Aspekte *Spezialisierung des Arztes*, *Prozedur* und *Medikation*, wird jeweils ein Sequenztyp gebildet. So wird beispielsweise für jede untersuchte Person eine Sequenz mit der Reihenfolge der Medikation erstellt, deren Symbole entweder die Ausprägung *ACE Hemmer* oder *Betablocker* haben. Die einzelnen Sequenztypen werden dann individuell geclustert. Hierzu wird das k-Medoids Clustering Verfahren mit der Editierdistanz *Longest Common Subsequence* (LCS) genutzt.

Insbesondere im Bereich der auf Abstandsmaßen basierenden Clusteringverfahren konnten bereits verschiedene Beispiele gefunden werden, die auf komplexeren Daten ausgeführt werden. In Herla et al. wird ein Clustering von Schneeprofilen beschrieben (42). Ein Schneeprofil besteht aus mehreren Schichten, die sich über die Zeit entwickeln. Jede Schneeschicht enthält Attribute, sowohl numerische als auch kategoriale, wie beispielsweise Schneekörnung und Schneehärte. Es werden verschiedene Abstandsmaße vorgestellt, um die einzelnen Attribute der Schneeschichten zu vergleichen. Hierauf aufbauend wird ein auf Dynamic Time Warping basierendes Clustering vorgestellt, bei dem die Attribute außerdem unterschiedlich gewichtet werden können. In Moreau et al. befassen sich die Autoren mit dem

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Clustern von multidimensionalen, semantischen Sequenzen (43). Als Anwendungsbeispiel dienen hier Tagesausflüge, die anhand der unternommenen Aktivitäten in Gruppen eingeteilt werden sollen. Eine Aktivität wird durch drei Attribute beschrieben: besuchter Ort, Art der Unternehmung und architektonischer Stil. Jedes Attribut ist ein Element einer Ontologie. Deren hierarchische Struktur wird zur Bestimmung des Abstandes von Attributen aus zwei unterschiedlichen Sequenzen herangezogen. Beispielsweise ähnelt der Besuch einer Bar dem Besuch eines Restaurants, da beide in der Ontologie als Orte, an denen etwas gegessen werden kann, aufgeführt werden. Zum Vergleich von zwei Sequenzen wird eine Editierdistanz verwendet. Die Autoren untersuchen unterschiedliche Clusteringverfahren und kombinieren diese zusätzlich mit dem Dimensionsreduktionsverfahren UMAP (Uniform manifold approximation and projection for dimension reduction) (44).

Die Ergebnisse der Recherche zum Clustering von sequentiellen Daten diente als Grundlage für die Konzeption und Implementierung der Data Mining Funktionalität im DMW, die im nächsten Abschnitt genauer beschrieben wird.

3.2.3 Konzeption

Anschließend wurde ein Konzept für das Clustering von klinischen Behandlungsverläufen erarbeitet. Da sich Krebsarten im Verhalten erheblich untereinander unterscheiden können und daher verschiedene Behandlungsmaßnahmen erfordern, ist das Clustering anhand von Behandlungsverläufen insbesondere innerhalb einzelner Krebsentitäten sinnvoll. Daher wurde der Fokus frühzeitig auf das Clustern von Brustkrebspatien*innen gelegt. Dabei wurde beschlossen, sich auf ähnlichkeitsbasierte sequentielle Clusteringverfahren zu fokussieren, da diese es besonders einfach machen, Expertenwissen in die Berechnung der Ähnlichkeit und somit in das Ergebnis des Clustering einfließen zu lassen.

Eine Möglichkeit, mit Sequenzen komplexer Daten, wie sie im SePaMiM Projekt vorliegen, umzugehen, besteht darin, die einzelnen Elemente zu kategorisieren und auf einfache Symbole abzubilden. Beispielsweise kann eine Operation im Kontext einer Brustkrebserkrankung anhand ihres OPS-Codes auf die Symbole *Brusterhaltene Therapie*, *Mastektomie* bzw. *sonstige Operation* abgebildet werden.

Ähnliche Abbildungen sind auch für andere Ereignistypen bzw. Krebsentitäten denkbar. Diese Sequenzen von Symbolen können dann im Anschluss geclustert werden. Als Abstandsmaße können hierzu beispielsweise Editierdistanzen genutzt werden. Ein solches Vorgehen wurde unter Verwendung der Levenshtein Editierdistanz und einem hierarchischen Clusteringverfahren als erster Proof-of-Concept umgesetzt. Bei der Diskussion der Ergebnisse wurde insbesondere festgestellt, dass bei der Abbildung auf einfache Symbole viele relevante Informationen verloren gehen.

Entsprechend wurde in einem nächsten Schritt ein Ähnlichkeitsmaß entwickelt, das sowohl die Ähnlichkeit einzelner Ereignisse als auch die Reihenfolge der Ereignisse berücksichtigt. Dabei werden Operationen, Strahlentherapien, systemischen Therapien und Krebsdiagnosen eines Patienten herangezogen. Die Ähnlichkeit zwischen verschiedenen Behandlungsereignissen wird anhand ausgewählter Attribute bestimmt. Bei Operationen sind dies zum Beispiel der OPS-Code, der Residualstatus und die Komplikationen. Für jedes Attribut wurde hierzu eine geeignete Distanzfunktion festgelegt. Das gewichtete Mittel der Abstände

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

einzelner Attribute ergibt dann den (normalisierten) Abstand zwischen zwei Ereignissen. Die berechneten Abstände zwischen einzelnen Ereignissen werden dann als Kosten einer Levenshtein Distanz genutzt.

Das beschriebene Ähnlichkeitsmaß wird dann dazu genutzt, mehrere Patient*innen paarweise miteinander zu vergleichen. Die hierbei entstehende Distanzmatrix wird im Anschluss mit dem UMAP-Dimensionsreduktionsverfahren auf 2 Dimensionen reduziert. Hierdurch können die Daten für einen besseren Überblick in einem Punktdiagram visualisiert werden. Außerdem erleichtert die Dimensionsreduktion das anschließende Clustering (45,46).

Der beschriebene Ansatz wurde auf einem Datensatz von 17822 Brustkrebspatient*innen aus dem LKR.NRW aus dem Diagnosejahr 2019 angewendet.

3.2.4 Evaluation

Für die Evaluation des Data Mining Ansatzes wurde ein Evaluationskonzept erstellt. Hier wurden drei allgemeine Evaluationsschritte definiert: ·

- die Überprüfung der Modellergebnisse anhand von technischen Kriterien·
- die inhaltliche Prüfung durch Fachexpert*innen des LKR
- die Untersuchung bezüglich klinisch relevanter Unterschiede zwischen den gefundenen Clustern.

Zu den technischen Kriterien gehörte die Überprüfung, ob sämtliche Patient*innen aus dem zu clusternden Datensatz im Ergebnis wieder aufzufinden sind und einem Cluster zugeordnet wurden. Weiterhin sollte die Heterogenität der Cluster, sowie die Homogenität innerhalb einzelner Cluster überprüft werden und die Anzahl der gefundenen Cluster beurteilt werden. Die inhaltliche Prüfung sollte zuerst eine rein deskriptive Auswertung der Clustereigenschaften umfassen und dann die eigentliche Prüfung der Expert*innen hinsichtlich der inhaltlichen Plausibilität innerhalb der Cluster und der inhaltlichen Unterschiede zwischen den Clustern beinhalten. Eine stichprobenartige Einzelfallanalyse, bei der sich Expert*innen einzelne Patient*innen anschauen und deren Clusterzuordnung bewerten, soll diesen Erfolgsparameter abrunden. Der dritte Erfolgsparameter soll den klinischen und fachlichen Nutzen der gefundenen Cluster bewerten. Hierbei sollen die Cluster anhand von klinisch relevanten Parametern, wie beispielsweise einer Überlebenszeitanalyse bewertet werden und Unterschiede hierin von Expert*innen bewertet werden.

Zur Überprüfung der technischen Plausibilität der genutzten Clustering-Verfahren wurde primär der Silhouette-Score genutzt. Allgemein ist es bei Cluster-Verfahren erstrebenswert, wenn die gebildeten Cluster möglichst homogen innerhalb der Cluster sind. Gleichzeitig sollten die Cluster untereinander möglichst heterogen sein. Der Silhouette-Score vereint die Homogenität innerhalb der Cluster und die Heterogenität zwischen den Clustern zu einem einzelnen Score und eignet sich also gut für eine Auswahl von Parametereinstellungen. Die inhaltliche Evaluation umfasste wie geplant eine deskriptive Auswertung einer Vielzahl von Clustern, sowie deren Bewertung in Bezug auf Zuordnung, Größe und Anzahl. Von einer detaillierten Einzelfallanalyse wurde abgesehen, da sehr oft bereits die vorhandenen Ergebnisübersichten die Gründe für Clusterbildung und Clusterzuordnung offenbarten.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Stattdessen wurde zur Evaluation des entwickelten Clustering-Tools eine qualitative Untersuchung mit Expertinnen und Experten der klinischen Auswertungsstelle des LKR NRW durchgeführt. Ziel war es, die Anwendbarkeit, Funktionalität und den Nutzen des Tools in der Praxis zu bewerten, insbesondere im Hinblick auf die Analyse von Patient*innenclustern. Sechs potenzielle Teilnehmende wurden kontaktiert, von denen drei nach einer Live-Demonstration bereit waren, an der Bewertung mitzuwirken. Die Datenerhebung erfolgte durch eine explorative Nutzung des Tools, wobei die Teilnehmenden selbst steuern konnten, welche Cluster sie näher untersuchen wollten. Zentrale Bewertungsaspekte waren die Konfiguration der Cluster, deren medizinische Plausibilität und mögliche Anwendungsfälle. Die gesammelten Rückmeldungen wurden qualitativ analysiert, um Stärken, Schwächen und Optimierungspotenziale des Tools zu identifizieren. Für die vollständige Methodik wird auf die Anlage 2, Evaluationsbericht Data Mining, verwiesen.

Zur Bewertung des klinischen und fachlichen Nutzens des Clusterings wurden zwei unterschiedliche Cluster-Konfigurationen untersucht: prognostisch günstige und prognostisch ungünstige Fälle. Für beide Konfigurationen wurde ein Clustering ausgeführt und im Anschluss eine Überlebenszeitanalyse, stratifiziert anhand der Clusterzuordnung, berechnet. Die Ergebnisse dieser Analysen wurden verglichen und diskutiert.

4 Projektergebnisse

In diesem Abschnitt werden die Evaluationsergebnisse des Pattern Matching Werkzeugs PMAW und des Data Mining Werkzeugs DMW kurz vorgestellt. Der Bezug zu den Forschungsfragen wird im Abschnitt 5 "Diskussion der Projektergebnisse" diskutiert. Zum Testen und Evaluieren der beiden Softwarewerkzeuge standen als Best-of aufbereitete Behandlungsverläufe von 1.058.706 Patient*innen mit insgesamt 7.688.816 Ereignissen zur Verfügung. Die jeweiligen Evaluationsberichte liegen als Anlage 1, Evaluationsbericht PMAW, und Anlage 2, Evaluationsbericht Data Mining, bei.

4.1 Projektergebnisse Pattern Matching

Ziel der Evaluation des PMAW war es, zu prüfen, ob sich die Anfragesprache CQL und das im Projekt entwickelte CQL Datenmodell für die Beantwortung von typischen Fragestellungen aus der klinischen Krebsregistrierung eignen. Hierfür wurden im Evaluationskonzept neun Fragestellungen der Krebsregister aus vier Bereichen definiert. Grundlegend war der Ansatz zu prüfen, ob sich die typischen Anfragen in CQL abbilden lassen und ob sich Ergebnisse der Anfragen mittels CQL von denen in den Krebsregistern verwendeten Anfragen unterscheiden (siehe auch Abschnitt 3.1.3 "Evaluation"). Die Ergebnisse der Umsetzung der Fragestellungen mit Hilfe von CQL werden im Folgenden kurz beschrieben. Eine umfangreiche Beschreibung dieser Ergebnisse wurde in Blohm et al. (47) veröffentlicht.

Bei der Implementierung der Qualitätsindikatoren QI6 und QI8 aus dem Bereich *Qualitätsindikatoren für Leitlinien* kam es zunächst zu Abweichungen zwischen den Ergebnissen der CQL Anfragen und den Ergebnissen der vorhandenen SQL Skripte des LKR.NRW. Die Abweichungen konnten darauf zurückgeführt werden, dass die von der Plattform § 65c, als Gremium zur Definition der Qualitätsindikatoren, beschriebenen Ein- und

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Ausschlusskriterien für die QIs in der Zwischenzeit aktualisiert wurden. Nach einer Korrektur der SQL Skripte im Krebsregister an die aktuelle Version der QIs wurden identische Ergebnisse erzielt.

Im Bereich *Plausibilitätsprüfung* wurden drei Beispiele für typische Plausibilitätsprüfungen aus dem LKR.NRW in CQL umgesetzt. Die Implementierung in CQL war dabei in allen drei Beispielen möglich. Bei der Plausibilitätsprüfung 2, zum Wechsel von palliativer zu kurativer Therapieintention, und bei Plausibilitätsprüfung 3, zum Rückgang der Tumorgröße ohne Therapie, gab es allerdings Abweichungen zwischen den Ergebnissen aus CQL und den vorhandenen SQL-Abfragen. Die Abweichungen waren jedoch nicht durch CQL begründet, sondern durch Unterschiede in der Datenmodellierung und -aufbereitung. Im CQL-Datenmodell werden beispielsweise bereits einige Qualitätsprobleme gelöst, die in den ursprünglichen Daten vorhanden waren. Um valide Zeitintervalle zu erzeugen, wird zum Beispiel das Start- und Enddatum von Behandlungsereignissen vertauscht, sollte das Startdatum nach dem Enddatum liegen. Diese Qualitätssicherung geschah im LKR.NRW zu dem Zeitpunkt noch nicht.

Die beiden Abfragen aus dem Bereich *Externe Datennutzung* konnten erfolgreich mit CQL implementiert werden. Die Ergebnisse sind identisch zu den vorhandenen Ergebnissen aus der KLast. Dabei wurden sowohl die Ein- und Ausschlusskriterien als auch die Abbildung der Daten in ein geeignetes Export Format in CQL umgesetzt.

Auch die beiden Abfragen für den interaktiven Onlinebericht der KLast, die dazu dienen, die Eignung von CQL für *Routineberichte* zu testen, konnten erfolgreich mithilfe von CQL implementiert werden. Dabei wurden die benötigten Ein- bzw. Ausschlusskriterien, die Eigenschaften für die Stratifizierung und die Projektion in ein JSON-Format für den Export in CQL implementiert. Auf Grundlage der exportierten JSON-Dateien wurden anschließend die benötigten Kennzahlen in R berechnet. Die Ergebnisse stimmten dabei mit vorhandenen Ergebnissen aus der KLast überein. Dieser Referenzergebnisse werden nicht in SQL erstellt, sondern mittels Anfragen in der Microsoft spezifischen Multidimensional Expressions (MDX) Anfragesprache auf einem speziell für multi-dimensionale Anfragen modellierten Microsoft SQL Server Data Warehouse.

Es konnte also zunächst rein technisch gezeigt werden, dass sich alle Repräsentanten für typische Fragestellungen der klinischen Krebsregister eignen. Damit war aus technischer Sicht bzgl. der Projektziele der Nachweis erbracht, dass sich oBDS-Datensätze, nach durchlaufen des Best-of-Prozesses, mit dem gewählten CQL Ansatz aufbereiten und analysieren lassen.

Die für die Evaluation beim LKR.NRW verwendeten Skripte und Projektentwicklungen sind beim OFFIS – Bereich Gesundheit auf Anfrage und bei berechtigtem Interesse zur Weiterverwendung unentgeltlich verfügbar. Bei den Anfragen im Rahmen der Evaluation der KLast zu Routineberichten sind die CQL Anfragen für PMAW und für die Externe Datennutzung auch die SQL Skripte enthalten. Die proprietäre Software zum Erstellen der Vergleichsdaten für die Routineberichte ist nicht verfügbar, dafür sind jedoch Beispielergebnisse enthalten.

Für alle Bereiche der Evaluation wurde jeweils ein Fallbeispiel in Anlage 3, PMAW CQL und SQL Vergleichs Anfragen der Evaluation, mit dem Hintergrund des Evaluationsberichts, der

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

CQL Anfrage für PMAW und den SQL Anfragen der Krebsregister für den „Goldstandard“ beigelegt.

4.2 Projektergebnisse Data Mining

Die Evaluation vom DMW war in drei Schritten unterteilt (siehe Abschnitt 3.2.4): eine technische und inhaltliche Prüfung sowie eine inhaltliche Bewertung.
Die Projektentwicklungen sind beim OFFIS – Bereich Gesundheit auf Anfrage und bei berechtigtem Interesse zur Weiterverwendung unentgeltlich verfügbar.

4.2.1 Technische Prüfung DMW

Für die technische Prüfung wurden die Silhouette-Scores der verschiedenen für DMW ausgewählten Clustering-Verfahren miteinander verglichen. Der Silhouette-Score kann Werte zwischen -1 und 1 annehmen, wobei ein höherer Wert ein besser separiertes Clustering bedeutet. Da wir für eine inhaltliche Auswertung keinen Anhaltspunkt zu einer sinnvollen Clusteranzahl hatten und es für sehr unterschiedliche Clusteranzahlen ähnlich gute Silhouette-Scores gab, haben wir die Clusteranzahl in drei Bereiche unterteilt, 1-50 Cluster, 51-100 Cluster und 101-200 Cluster. In jedem Bereich wählten wir unter den Clusteringverfahren DBScan, HDBScan und Hierarchisches Clustern mit verschiedenen Parametereinstellungen dasjenige mit dem höchsten Silhouette-Score aus. Außerdem haben wir die Anzahl von als Ausreißern gekennzeichneten Verläufe auf maximal 5% der Gesamtdaten beschränkt. Der UMAP-Parameter n_neighbors wurde auf 25 festgelegt. Andere Parameter wurden auf ihrem Standardwert belassen. In allen drei Fällen hatte das hierarchische Clusterverfahren die besten Silhouette-Scores. Deshalb wurde für die weitere inhaltliche Prüfung und Bewertung die Auswertungen mittels des hierarchischen Clusterns verwendet.

4.2.2 Inhaltliche Prüfung DMW

Im Folgenden werden einige der gefundenen Cluster beispielhaft beschrieben, die inhaltlich geprüft und bewertet wurden. Weitere Details und Beispiele sind in Anlage 2, Evaluationsbericht Data Mining, wiedergegeben.

Die Abbildung 1 zeigt die UMAP-Projektion eingefärbt in den Clustern der Methode mit dem höchsten Distance Threshold (TSH) = 15 und der geringsten Clusterzahl. Es gibt zwölf von der Größe her sehr unterschiedliche Cluster (Nummeriert von 0 - 11), die größtenteils auch optisch in der UMAP-Projektion den dort gebildeten Punktewolken entsprechen.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

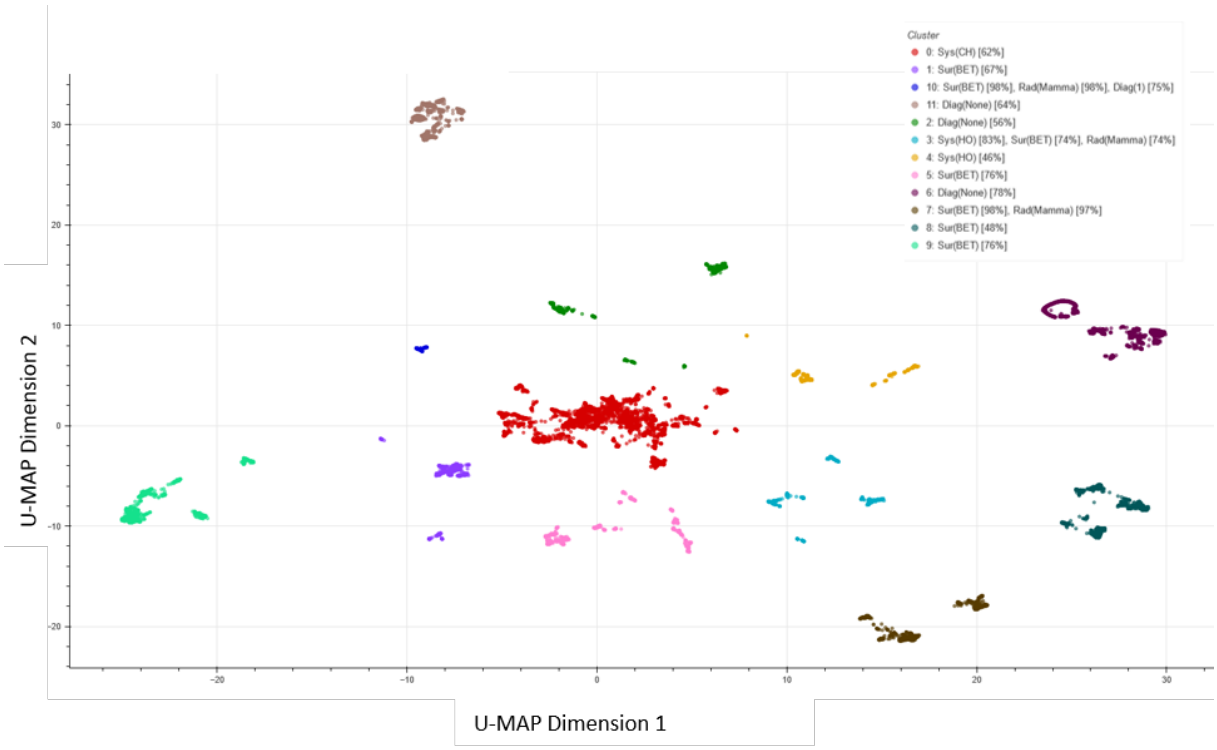


Abbildung 1: Cluster nach UMAP-Projektion mit TSH=15

Ebenfalls sind in der Abbildung 1 rechts oben die häufigsten Ereignisse zu sehen, die diese Gruppe kennzeichnen, sowie die Angabe, wie viele der Fälle im Cluster diese Ereignisse mindestens einmal enthalten.

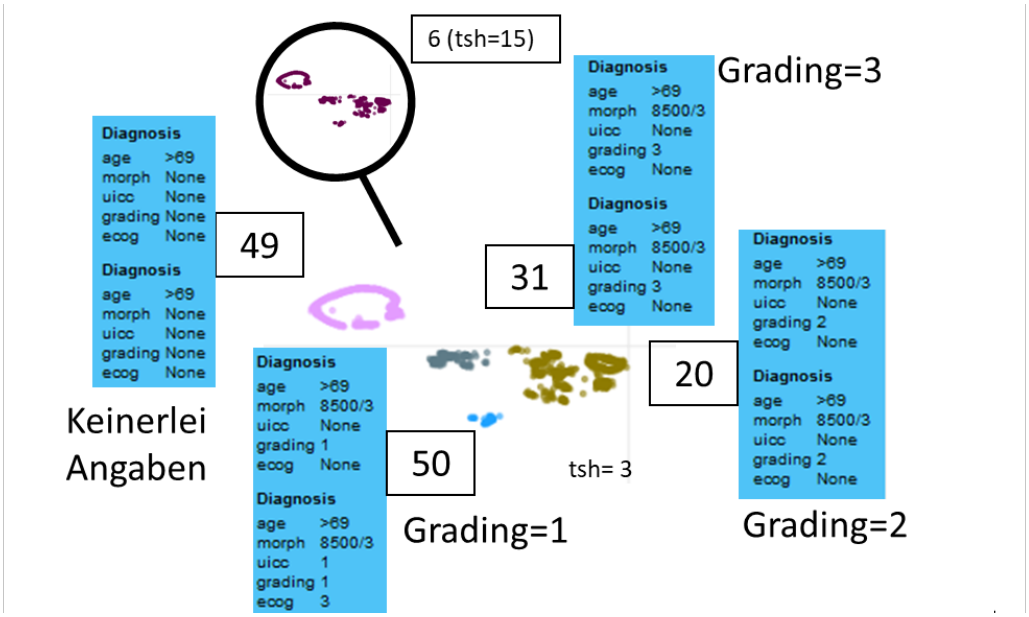


Abbildung 2: Darstellung von Cluster-Details

Oft war es gut erkennbar anhand welcher Kriterien die Fälle auf die Cluster verteilt wurden – manchmal schon in der geringsten Auflösung mit 12 Clustern, oder auch erst in einer der beiden höheren Auflösungen (TSH=3 oder 1). So enthält das Cluster 6 (TSH=15) die Fälle mit

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Diagnosealter >69 ohne weitere Behandlung (siehe Abbildung 2) und wird in der nächst höheren Auflösung in mehrere Cluster (TSH=3, Cluster 20, 31, 49 und 50) aufgeteilt. In dieser Altersgruppe scheint das Grading für die weitere Aufteilung ausschlaggebend zu sein, was aus fachlicher Sicht sinnhaft erscheint.

In Abbildung 3 ist die Behandlungssequenzansicht des Clusters 21 aus der mittleren Auflösung (TSH=3) zu sehen. Von links nach rechts sieht man die Art des Ereignisses in chronologischer Reihenfolge beginnend mit dem ersten (hier: Diagnose). Jeder Zeile entspricht einem Brustkrebs-Fall. In der Legende (rechts neben der Behandlungssequenzansicht) sieht man die Eigenschaften, die ein solches Ereignis haben kann. Daraus ist ersichtlich, dass in dem visualisierten Cluster hauptsächlich Fälle mit Operation, gefolgt von Chemotherapie und in einigen Fällen auch noch einer Radiotherapie zusammengefasst sind.

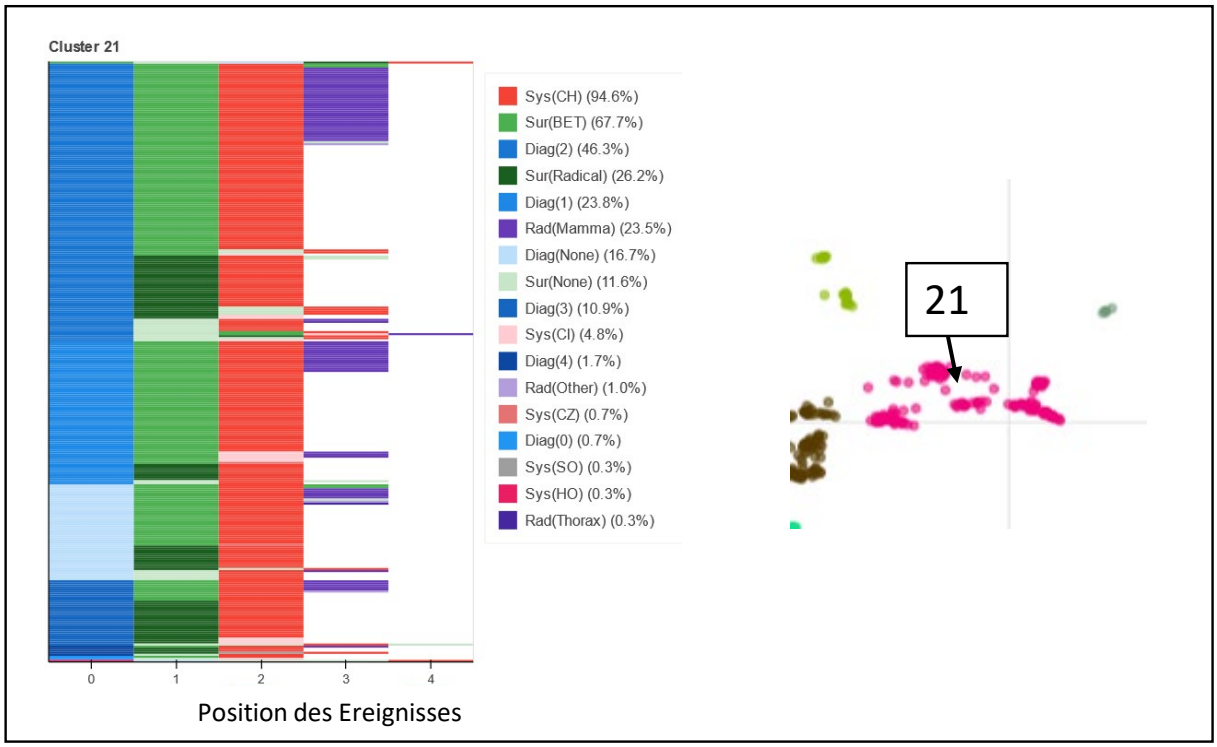


Abbildung 3: Darstellung von Cluster 21 in der U-MAP-Projektion (TSH 3, rechts) und dessen Behandlungssequenzansicht.

In der noch feineren Auflösung wird Cluster 21 (TSH=3) anhand des Behandlungsmusters auf drei Cluster aufgeteilt (TSH=1 Cluster 11, 50 und 144, siehe Abbildung 4). Das Modell mit TSH=1 sammelt in Cluster 50 und 144 nur Fälle mit brusterhaltender Therapie und Chemotherapie, in 144 ist zusätzlich noch eine Bestrahlung dabei. Cluster 11 hingegen enthält mehrheitlich eine radikale OP und Chemotherapie, jedoch keine Bestrahlung. In allen drei Clustern, aber vermehrt in Cluster 11 finden sich interessante Ausreißer, die mehr Ereignisse als die anderen Fälle haben oder bei denen die Reihenfolge vertauscht ist. So ist in Cluster 11 z.B. ein Fall mit systemischer Therapie vor Diagnose zu finden, was implausibel ist.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

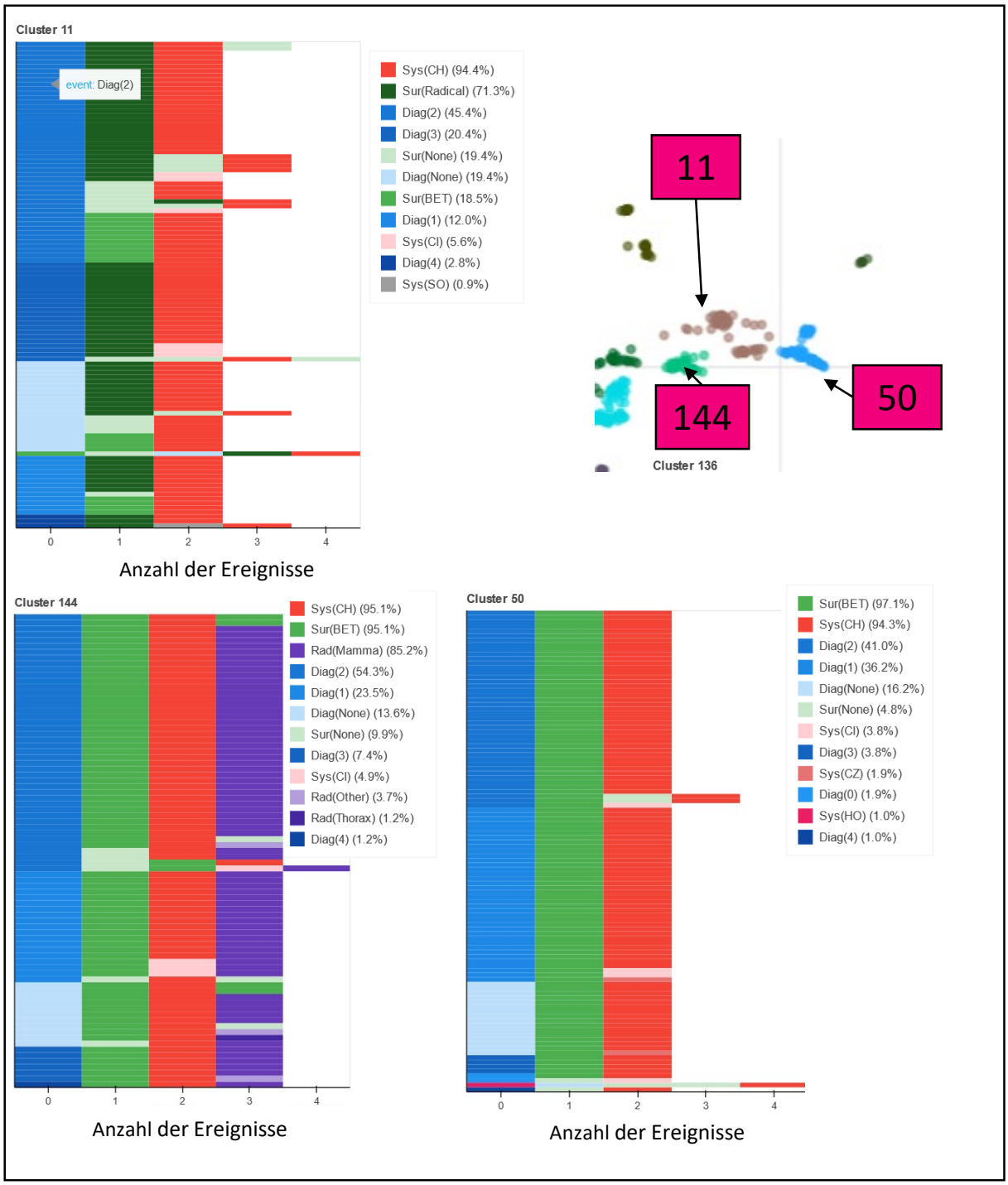


Abbildung 4: Darstellung der Cluster 11, 144, 50 in der U-MAP-Projektion (TSH 1, rechts oben) und deren Behandlungssequenzansichten

Das Clustering-Verfahren konnte uns außerdem mehrere Cluster mit offenbar implausiblen Daten zeigen. Darunter befindet sich ein Cluster mit unter 50 Jährigen ohne eine Behandlung. Es ist äußerst unwahrscheinlich, dass Patient*innen unter 50 Jahren nicht in irgendeiner Form behandelt werden. Es ist zu vermuten, dass bei diesen Fällen die entsprechenden Behandlungsdaten nicht ans Krebsregister übermittelt wurden.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Im Expert*innen-Interview wurden die Clustering Ergebnisse mit allen drei Thresholds, sowie deren Darstellung besprochen. Die UMAP-Projektion wurde als sehr gut geeignet für die visuelle Darstellung des Clustering-Ergebnisses empfunden. Als geeignet für weitere inhaltliche Analysen wurde das Model mit 53 Clustern (TSH=3) befunden. Im Verfahren mit wenigen, gröberen Clustern könnte man sich ein interessantes Cluster heraussuchen und dieses in den Verfahren mit der höheren Clusteranzahl suchen, um so eine feinere Aufteilung der Behandlungsverläufe zu analysieren.

4.2.3 Plausibilität DMW

Zur Beurteilung der Plausibilität der Cluster wurde bei der Sichtung der verschiedenen hierarchischen Cluster Dashboards in viele Cluster genauer geschaut (per Zoomen und Tooltip) und bei den meisten sowohl die Abgrenzung als Cluster als auch die Homogenität innerhalb der Cluster als plausibel befunden.

Es wurden mehrere Vorschläge geäußert, wie das Clustering der Behandlungsabläufe in der Zukunft in der Praxis angewendet werden könnte. Eine dieser Ideen war, Clustering zur Überprüfung der Datenqualität zu nutzen, indem auffällige oder unerklärliche Cluster analysiert werden. Z. B. wurden Cluster mit Therapien vor Diagnose oder brusterhaltende Therapie nach Mastektomie beobachtet.

Abschließend wurden mögliche Anpassungswünsche an das Modell diskutiert, um es für den Einsatz in der Praxis zu optimieren. Dies waren z.B. eine stärkere Datenbereinigung, die Berücksichtigung zeitlicher Abstände zwischen Ereignissen, die Unterscheidung von Erst- und Zweitlinientherapie sowie die Einbeziehung von Multimorbidität und Tod. Auch das Clustern auf Patient*innenebene und eine veränderte Gewichtung zwischen Diagnose und Behandlung wurden vorgeschlagen.

4.2.4 Klinisch-fachlicher Nutzen DMW

Ein weiterer Aspekt zur Bewertung der Modelle war die Frage, ob die gefundenen Cluster klinisch relevante Unterschiede in den Behandlungsverläufen zeigen. Häufig wird bei Krebsregisterdaten eine Überlebenszeitanalyse durchgeführt. Dafür kommen unter anderem der Kaplan-Meier-Ansatz oder das relative Überleben als Methoden zur Anwendung. Der klinisch-fachliche Nutzen war dadurch definiert, dass die durch das Clustering identifizierten Behandlungsverläufe einen prognostischen Wert haben, der sich durch Unterschiede im 4-Jahres Kaplan-Meier-Überleben zwischen den Clustern äußert, selbst wenn für mögliche Verzerrungen durch systematische Unterschiede in Diagnosealter oder bestimmten Tumoreigenschaften zwischen den Clustern adjustiert wird. Für die Bewertung des klinisch-fachlichen Nutzens wurden zwei Patient*innenkohorten (prognostisch ungünstige (PU) Fälle und prognostisch günstige (PG) Fälle) geclustert und die einzelnen Cluster der beiden Kohorten mit anhand einer Überlebenszeitanalyse bewertet.

Zusammenfassend zeigt sich, dass das Clustering - auch ohne Berücksichtigung des Überlebens bei der Berechnung der Abstandsmaße der Fälle - prognostisch deutlich abgrenzbare Gruppen von Therapieverläufen findet, die sich nicht allein durch unterschiedliche Verteilungen der Baseline-Charakteristika der Patientinnen erklären lassen. Der Ansatz hat das Potential die Krebsregisterdaten explorativ auszuwerten und

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Auswerter*innen auf unerwartete Zusammenhänge von Überlebenswahrscheinlichkeit und Therapieverlauf aufmerksam zu machen. Diese Ergebnisse können unter anderem dabei helfen, neue Forschungshypothesen zu generieren und mangelnde Daten- und Versorgungsqualität aufzudecken. Die Identifikation potentieller Risikogruppen in der Krebsversorgung kann durch einen Transfer in die wissenschaftliche und klinische Praxis zu einer Verbesserung der integrativen Krebsbehandlung beitragen.

5 Diskussion der Projektergebnisse

In diesem Abschnitt werden zunächst die Projektergebnisse der beiden Projektteile PMAW und DMW diskutiert. Im Anschluss wird Bezug darauf genommen, inwiefern die im Projekt gestellten Forschungsfragen beantwortet werden konnten.

5.1 Diskussion Pattern Matching

Im Rahmen des Projektes konnte gezeigt werden, dass CQL dazu geeignet ist, die Krebsregister bei der Umsetzung verschiedener Routineaufgabe zu unterstützen (47). Insbesondere ist es mit CQL möglich, komplexe Ein- und Ausschlusskriterien anhand von Behandlungsverläufen zu definieren. Hierunter fällt beispielsweise das Filtern von Behandlungen anhand verschiedener Attribute, aber auch die Formulierung von zeitlichen Beziehungen zwischen mehreren Behandlungen. Ein wichtiger Aspekt ist hierbei neben der Formulierung der Anfragen auch deren Lesbarkeit. Nach Meinung der Mitarbeitenden der am Projekt beteiligten Krebsregister waren CQL-Anfragen im Allgemeinen leichter verständlich als vorhandene SQL-Anfragen. Dies kann dazu beitragen, die Diskussion über oftmals komplexe Kriterien zur Definition einer Kohorte zu erleichtern. Die Nutzerfreundlichkeit bzw. leichtere Anwendung durch nicht technisches Fachpersonal wurde nicht durch eine Nutzerstudie evaluiert, da die technischen Aspekte von CQL im Fokus der konkreten Evaluation standen.

Die Verwendung von CQL als eine einheitliche Abfragesprache auf einem einheitlichen, für Analysezwecke aufbereiteten Datenmodell, könnte die Kommunikation und gemeinsame Nutzung von Abfragen zwischen Krebsregistern erleichtern und zur Standardisierung verschiedener Auswertungen beitragen. Beispiele hierfür sind die Online Berichte oder die Qualitätsindikatoren. Hier gibt es das Ziel, möglichst vergleichbare Auswertungen zu erstellen. Die Verwendung einer einheitlichen, formaleren Abfragesprache kann es ermöglichen, unterschiedliche Interpretationen und somit inkonsistente Implementierungen zu vermeiden.

Die Verwendung von CQL hat jedoch auch Nachteile. CQL ist als domänenspezifische Sprache weniger bekannt als gängige Programmiersprache wie R, Python oder SQL. Entsprechend ist oftmals eine Einarbeitung erforderlich. Auch hat CQL mit dem Fokus auf klinische Daten einen geringeren Funktionsumfang als gängige Programmiersprachen. CQL ermöglicht zwar die einfache Definition von Kohorten, eine Weiterverarbeitung der Daten muss allerdings unter Umständen in externen Werkzeugen erfolgen.

Ein entsprechend wichtiger Aspekt bei der Implementierung des PMAW war deshalb auch die Visualisierung der Ergebnisse einer CQL Abfrage. Neben einfachen deskriptiven Statistiken zu einer Kohorte werden hier insbesondere die Behandlungsverläufe visualisiert. Dies geschieht

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

in aggregierter Form für die ganze Kohorte mit Hilfe eines Sankey Diagramms und für individuelle Patient*innen in Form von interaktiven Zeitreihen.

Die grafische Oberfläche des entwickelten Demonstrators PMAW ist in Abbildung 5 zu sehen. Die Oberfläche bietet die Funktionalität zur Definition und Eingabe der CQL Anfragen und eine tiefergehende Visualisierung der Ergebnisse. So ist es nicht nur möglich, sich die gebildeten Kohorten in Form von Identifikatoren-Listen für weiter Anfragen geben zu lassen, sondern erfährt auch grundlegende Informationen zu Durchschnittswerten und Verteilungen. Es können auch Detailinformationen jedes Behandlungsfalls in der Kohorte einzeln angezeigt werden.

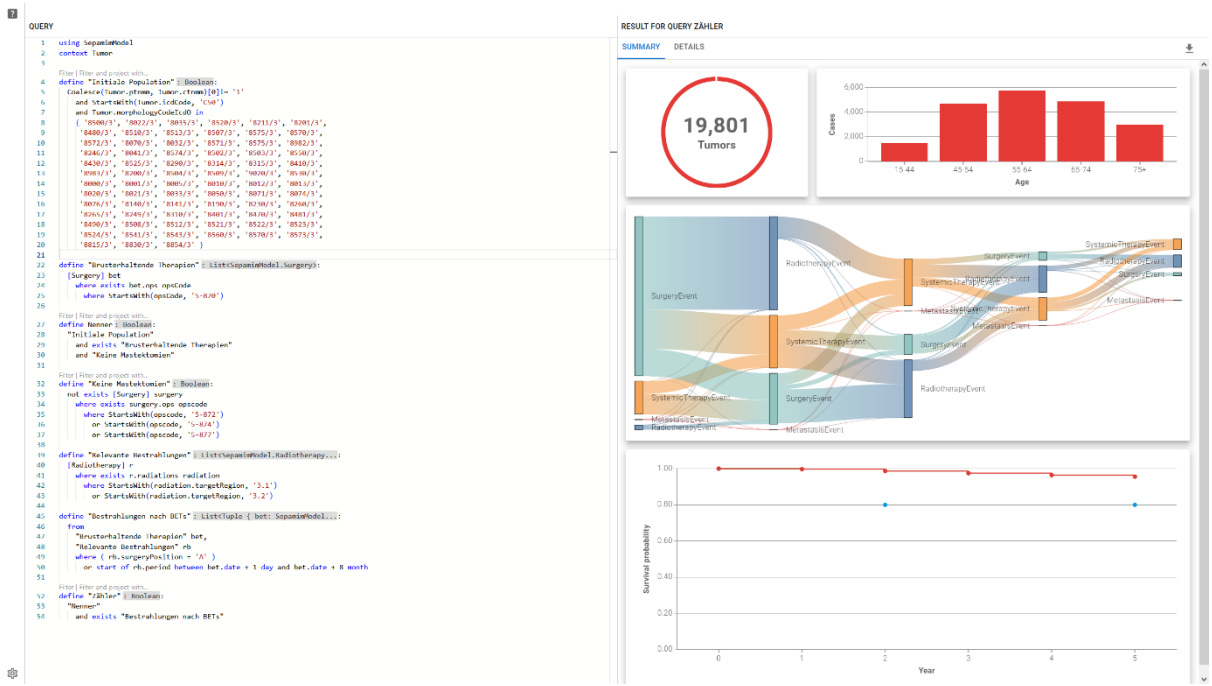


Abbildung 5: Darstellung PMAW mit CQL- Definition auf der linken und der Ergebniskohorte auf der rechten Seite. Der gezeigte Datenstand ist fiktiv.

Der Demonstrator hat eine voll funktionsfähige rudimentäre Eingabemaske und bereits umfangreiche Ausgabenvisualisierung. Die Konfiguration findet mittels textbasierter Konfigurationsdateien statt. Das darunterliegende Informationssystem benötigt aus den oBDS-Rohmeldungen abgeleitete Best-of-Daten. Da der oBDS bundesweit standardisiert ist, sind in allen klinischen Krebsregistern die entsprechenden Informationen vorhanden. Da die technische Umsetzung des Best-of aber Registerindividuell ist (auch in der Datenrepräsentation) muss hier jeweils ein geeignetes Mapping stattfinden. Es wurden Zuordnungen einzelner Ereignisse zu Ereignisklassen im Projekt mit dem LKR.NRW und KLAst definiert, die von anderen Krebsregister genutzt werden können. Dieses Wissen muss derzeit manuell in die Anfrage codiert werden.

Die Anwendungen laufen stabil und sind getestet. Aus Entwicklersicht müssten Sie für einen Produktivbetrieb aber mit einer weiterentwickelten Eingabemaske, die auch Hilfestellungen gibt und Fehlbedienungen verhindert, ausgestattet werden. Zudem wurde bei der Entwicklung nur die Nutzeranforderungen von zwei Krebsregistern von insgesamt 16 (15 Bundesländer + ZfKD) eingeholt.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Der klinische Nutzen vom PMAW ergibt sich aus den Aufgaben der klinischen Krebsregistrierung, die über Qualitätsindikatoren die Versorgungsqualität an die Melder zurückspiegelt mit dem Ziel, dass diese ihre Qualität verbessern. Zudem werden Daten für externe (klinische) Forschungsanfragen und die Versorgungsforschung bereitgestellt. Damit lässt sich z.B., im Gegensatz zu klinischer Versorgung, die Effektivität von Therapien in einem Real-World-Setting vergleichen. Die Routineberichterstattung dient der Politikberatung sowie Gesundheitsplanung und ist damit eher von informierender, strategischer Bedeutung. PMAW unterstützt die Krebsregister bei allen diesen Aufgaben und erleichtert diese.

5.2 Diskussion Data Mining

Da bei vielen Clustern die Diagnosefaktoren (z.B. Alter und Grading) eine große Rolle spielen, könnte man bei einer Wiederholung des Verfahrens ins Betracht ziehen von vorneherein eine homogene Ausgangssituation (z.B. Tumore mit niedrigem oder hohem Risiko, etc.) herzustellen und die Diagnose nicht für das Clustering zu benutzen, um den Fokus mehr auf die Behandlungsverläufe zu legen Andererseits zeigt aber auch das hier evaluierte Modell bereits sehr viele sich klar voneinander abgrenzenden Cluster mit konkreten Behandlungsverläufen, darunter sowohl bekannte und den Leitlinien entsprechende als auch eher ungewöhnliche oder welche, die auf mangelnde Datenqualität hinweisen.

All diese Behandlungseinteilungen sind durchaus sinnvoll und zeigen, dass die Grundidee, ein Clusterverfahren dazu zu benutzen, erfolgreich war.

Bzgl. des Überlebens identifiziert das Clustering klar abgrenzbare Gruppen von Therapieverläufen, die nicht allein durch die Baseline-Charakteristika der Patientinnen erklärt werden können. Dieser Ansatz ermöglicht eine explorative Auswertung von Krebsregisterdaten und kann auf unerwartete Zusammenhänge zwischen Überlebenswahrscheinlichkeit und Therapieverlauf hinweisen, was neue Forschungshypothesen generieren und Qualitätsmängel aufdecken könnte. Wichtig ist in diesem Zusammenhang zu erwähnen, dass es in dieser Analyse zu Verzerrungen durch nicht bekannte Faktoren, wie Komorbiditäten oder weitere Tumoreigenschaften gekommen sein kann, die sowohl mit der gewählten Therapie als auch dem Überleben zusammenhängen. Die gefundenen Überlebensunterschiede dürfen also nicht als Therapieempfehlungen verstanden werden.

DMW wurde, anders als PMAW, nicht als Anwenderwerkzeug entwickelt, sondern als Spezialsoftware, die von IT-Expert*innen / Data Scientists bedient wird. Dies geschah insbesondere im Hinblick darauf, dass lediglich die technischen Möglichkeiten und inhaltliche Sinnhaftigkeit von Data Mining evaluiert werden sollte. Die Umsetzung erfolgt mittels spezieller Skriptsprachen und wurden als Skript-Sammlungen (Jupiter Notebooks) gespeichert. Notwendige Parametrisierungen und Feinjustierungen geschahen in dieser Softwareumgebung. Es gibt also keine Benutzeroberfläche oder andere Konfigurationsdateien, welche Nicht-IT-Expert*innen im Krebsregister dazu befähigen würden das Clustering an ihre Gegebenheiten anzupassen. Die Visualisierung der Ergebnisse, welche bereits umfangreich sind und sehr gute Einblicke in die Cluster geben, stellen die

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Anwendersicht dar und wurden von den Expert*innen der Krebsregister als sehr zielführend wahrgenommen.

Da die Festlegung des Abstandsmaßes für jede Krebsentität einzeln erfolgen muss und in SePaMiM beispielhaft für Brustkrebs implementiert wurde, ist eine Übertragung des Modells auf andere Krebs-Entitäten oder Behandlungsschwerpunkte nicht ohne eine zeitintensive Überarbeitung von Expert*innen möglich. Jedes Abstandsmaß muss manuell und individuell an die jeweilige Fragestellung angepasst werden, z.B. wurde beim Mammakarzinom festgelegt, dass bei den Operationen nur zwischen brusterhaltender Therapie, Mastektomie und sonstigen Operationen unterschieden werden soll.

5.3 Bezug zu den Forschungsfragen

Die primäre Forschungsfrage des SePaMiM Projektes lautet: „Wie lassen sich die im onkologischen Basisdatensatz enthaltenen Behandlungsverläufe aufbereiten und analysieren, so dass sie zur Unterstützung der Versorgungsforschung in klinischen Krebsregistern genutzt werden können?“. Diese wird durch die folgenden zwei Arbeitshypothesen genauer spezifiziert:

- 1. Durch ein geeignetes Datenmodell und die Verwendung von komplexen Suchverfahren können Expert*innen gezielt Kohorten mit unterschiedlichen Behandlungsmustern bilden und diese anhand gängiger Analyseverfahren, wie z.B. Überlebenszeitanalysen, miteinander vergleichen, woraus sich Hinweise auf potentiell geeignetere Behandlungspfade ergeben können.
- 2. Die Verwendung von KI und Data Mining Verfahren bietet die Möglichkeit, Patient*innenkohorten zu identifizieren, als Voraussetzung für die Suche nach Ursachen von Versorgungsunterschieden

Die erste Arbeitshypothese wurde im Rahmen des Pattern Matching Ansatz bearbeitet. In der Evaluation konnte gezeigt werden, dass CQL zusammen mit einem im Projekt entwickelten Datenmodell dazu geeignet sind, die Krebsregister bei der Bildung von Kohorten für unterschiedliche Fragestellungen zu unterstützen.

Bei den verschiedenen Clusterverfahren wurden sowohl Cluster mit bereits bekannten Behandlungsverläufen gebildet, deren Zuordnung zu einem Cluster direkt nachvollziehbar war, als auch unerwartete Cluster, die die Expert*innen bei einer manuellen Suche nicht definiert hätten. Daher ist es bei entsprechender Parameterauswahl hervorragend dazu geeignet alternative Patient*innenkohorten zu identifizieren, die eine hohe Ähnlichkeit untereinander besitzen und anders vielleicht nicht erkannt worden wären.

Zudem wurden vier sekundäre Forschungsfragen definiert:

- (2.1) Lassen sich aus den gewonnenen Erkenntnissen Standards für Routineauswertungen in den Krebsregistern etablieren (z.B. Abgleich mit leitliniengerechten Therapiealgorithmen, Validierung von Datensätzen hinsichtlich Vollständigkeit des Behandlungsverlaufs)?
- (2.2) Lassen sich alternative Behandlungs- und Erkrankungsverläufe erkennen / Unterschiede zwischen Krankenhäusern / Regionen, etc.?

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

- (2.3) Wie lassen sich so gefundene Kohorten sinnvoll analysieren? Z.B. Überlebenszeitanalysen
- (2.4) Lässt sich die Qualität der Daten bewerten?

In Bezug auf die erste sekundäre Forschungsfrage 2.1 kann CQL in Kombination mit einem einheitlichen Datenmodell zur Standardisierung von verschiedenen Auswertungen beitragen. In den USA findet CQL hierzu bereits Anwendung (3,48–50).

Die zweite sekundäre Forschungsfrage 2.2 wurde sowohl vom Pattern Matching- als auch vom Data Mining Ansatz adressiert. Mit Hilfe des PMAW lassen sich Kohorten für alternative Behandlungspfade gezielt abfragen und im Anschluss analysieren. Der Data Mining Ansatz kann dazu genutzt werden, automatisiert Patient*innen in Kohorten einzuteilen und somit alternative Behandlungspfade zu finden.

Sowohl im PMAW, wie auch im DMW wurden unterschiedliche Visualisierungen und Ansichten zur weiteren Analyse der gefundenen Kohorten implementiert. Der Fokus lag hierbei insbesondere auf der Visualisierung der Behandlungspfade. Hierbei wurden sowohl Behandlungspfade einzelner Tumore als auch aggregierte Übersichten von mehreren Behandlungspfaden einer Kohorte umgesetzt. Diese Arbeiten dienten der Beantwortung der dritten sekundären Forschungsfrage 2.3.

Wie in der Evaluation des PMAW gezeigt wurde, ist es mit diesem möglich, Plausibilitätsprüfungen durchzuführen, um die Datenqualität zu verbessern. Bei der Expert*innendiskussion der Clusteringergebnisse aus DMW wurden u.a Cluster gefunden, die sich auf Datenqualitätsprobleme zurückführen ließen. Beide Ansätze tragen somit zur Beantwortung der vierten sekundären Forschungsfrage 2.4 bei.

5.4 Limitationen des entwickelten Vorgehens und der Werkzeuge

Bzgl. der in der Projektlaufzeit entwickelten Vorgehen und Werkzeuge gibt es Einschränkungen. Die wichtigste ist, dass die Werkzeuge nicht direkt auf den einzelnen oBDS-Meldungen arbeiten, sondern Analysen erst auf dem aufbereiten Best-of-Datensatz durchgeführt werden können. Dadurch sind die gezeigten und gemachten Auswertungen zwar prinzipiell für alle deutschen Krebsregister durchführbar. Jedoch muss technisch für jedes Krebsregister einmalig ein Mapping zwischen dem Schema des jeweiligen Best-of-Datensatz zur in SePaMiM genutzten Struktur erfolgen. Alle Inhalte sind aber aufgrund des standardisierten oBDS-Datensatzes verfügbar.

Für DMW besteht eine Limitation darin, dass es keine Benutzerobfläache gibt und ein exploratives Ausprobieren, welches vielen Data Mining Verfahren inherent ist, nicht ohne Programmierkenntnisse möglich ist.

Außerdem kann jede Änderung des Datensatzes oder der Parameter das Ergebnis beeinflussen und erschwert daher die Reproduzierbarkeit in anderen Krebsregistern oder mit anderen Entitäten. Nicht zuletzt limitiert die Datenqualität auch die Qualität des Ergebnisses. Wie in Abschnitt 4.2.2 gezeigt erzeugen vermutlich fehlende Therapiedaten eigene Cluster

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

und so ist auch nicht auszuschließen, dass Fehlen einzelner Merkmale die Clusterbildung beeinflusst.

6 Verwendung der Ergebnisse nach Ende der Förderung

Im Laufe des Projektes wurden die Projektergebnisse regelmäßig zu unterschiedlichen Anlässen relevanten Interessensgruppen präsentiert. Dies geschah beispielsweise im Rahmen des von OFFIS organisierten CARESS Anwender-Workshops, an dem ca. 60 Expert*innen aus dem Bereich der Krebsregistrierung teilnahmen, im Kontext der vom BMG initiierten AG „Vergleichbare Erkrankungen“ der Plattform § 65c Gruppe, auf der 68. und 69. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) in Heilbronn und Dresden, sowie in mehreren Treffen mit einzelnen Landeskrebsregistern. Insgesamt wurde großes Interesse an den Ergebnissen des Projekts bekundet, so dass geplant ist, diese bei den Krebsregistern zum Einsatz zu bringen.

Weiterer Forschungsbedarf besteht insbesondere noch beim Data Mining Ansatz, wo geplant ist, weitere Abstandsmaße zu evaluieren und das Clustern weiterer Diagnosegruppen zu unterstützen. Weiterhin sind noch Entwicklungsarbeiten notwendig, um die bestehenden Forschungsprototypen einem produktiven Einsatz zuzuführen. Beispielsweise soll das CQL-Datenmodell weiterentwickelt werden, wozu weitere Krebsregister einbezogen werden sollen, um dem Ziel eines einheitlichen Datenmodells näher zu kommen.

Für einen bundesweiten Einsatz der Anwendungen braucht es entweder eine Standardisierung innerhalb der Krebsregister oder jeweils angepassten Datenaufbereitungen. Zudem müssten aus den Prototypen nutzerfreundliche Endanwendungen entwickelt werden.

7 Erfolgte bzw. geplante Veröffentlichungen

Tabelle 2: Übersicht erfolgter bzw. geplanter wiss. Veröffentlichungen

Titel	Veröffentlichungsstatus
SePaMiM – an online tool for analyzing course-of-disease data in German cancer registries using CQL (Abstract) Kolja Blohm, David Korfkamp, Christian Lüpkes, Andreas Hein; 68. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS); 2023 https://dx.doi.org/10.3205/23gmds067	Veröffentlicht Vortragsabstract
Clustering breast cancer patients based on their course of treatment (Abstract) Kolja Blohm, David Korfkamp, Christian Lüpkes; 69. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS); 2024doi: https://dx.doi.org/10.3205/24gmds137	Veröffentlicht Vortragsabstract
The Clinical Quality Language as a tool to support data analysis in German clinical cancer registries.	Veröffentlicht (Peer Reviewed Research Article)

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Kolja Blohm, David Korfkamp, Joachim Hübner, Florian Oesterling, Stefanie Schulze, Andreas Hein. GMS Med Inform Biom Epidemiol. 2024;20:Doc11. https://dx.doi.org/10.3205/mibe000267	
Clustering Treatment Courses of Breast Cancer Patients Using German Cancer Registry Data (Full Paper) Kolja Blohm, David Korfkamp, Florian Oesterling, Stefanie Schulze, Andreas Hein.	Geplant

IV Literaturverzeichnis

Da das Dokument keine Fußnoten zulässt, wurden im Literaturverzeichnis URLs als Quellenangaben zu Softwarebibliotheken, Informationssystemen etc. hinterlegt.

1. Stegmaier C, Hentschel S, Hofstädter F, Katalinic A, Tillack A, Klinkhammer-Schalke M. Das Manual der Krebsregistrierung. 2019.
2. Arbeitsgemeinschaft Deutscher Tumorzentren e.V. (ADT). Einheitlicher onkologischer Basisdatensatz. 2021 [cited 2024 Dec 18]. Bundeseinheitlicher Onkologischer Basisdatensatz. Available from: https://www.basisdatensatz.de/basisdatensatz
3. Health Level Seven International. Cross-Paradigm Specification: Clinical Quality Language, Release 1. 2020 [cited 2024 Mar 14]. Clinical Quality Language (CQL). Available from: https://cql.hl7.org/
4. Vom Brocke J, Simons A, Niehaves B, Niehaves B, Riemer K, Brocke J, et al. RECONSTRUCTING THE GIANT: ON THE IMPORTANCE OF RIGOUR IN DOCUMENTING THE LITERATURE SEARCH PROCESS. In: ECIS 2009 Proceedings 161 [Internet]. 2009. Available from: https://aisel.aisnet.org/ecis2009/161
5. International Organization for Standardization (ISO). Standard, International Organization for Standardization. 2016 [cited 2016 Jun 15]. Information technology — Database languages — SQL Technical Reports — Part 5: Row Pattern Recognition in SQL. Available from: https://www.iso.org/standard/65143.html
6. Buchmann A, Koldehofe B. Complex Event Processing. itit. 2009 Sep;51(5):241–2.
7. Diao Y, Immerman N, Gyllstrom D. SASE+: An Agile Language for Kleene Closure over Event Streams [Internet]. 2007 [cited 2024 Dec 18]. Available from: https://www.lix.polytechnique.fr/~yanlei.diao/publications/07-03.pdf
8. Demers A, Gehrke J, Panda B, Riedewald M, Sharma V, White W. Cayuga: A General Purpose Event Monitoring System [Internet]. 2007. Available from: http://www.complexevents.com,
9. Dindar N, Güç B, Lau P, Ozal A, Soner M, Tatbul N. DejaVu. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. New York, NY, USA: ACM; 2009. p. 1023–6.
10. Holford M. Neo4j Developer Blog. 2020 [cited 2024 Apr 16]. Modeling Patient Journeys with Neo4j. Available from: https://medium.com/neo4j/modeling-patient-journeys-with-neo4j-d0785fbbf5a2
11. Neo4j Inc. Neo4j [Internet]. 2024 [cited 2024 Dec 18]. Available from: https://neo4j.com/

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

12. Bitnine Inc. AgensGraph Quick Guide [Internet]. 2016 [cited 2024 Dec 18]. Available from: https://github.com/bitnine-oss/agensgraph?tab=readme-ov-file
13. ArangoDB GmbH. ArangoDB [Internet]. 2024 [cited 2024 Dec 18]. Available from: https://arangodb.com/
14. JanusGraph Authors. JanusGraph [Internet]. 2024 [cited 2024 Dec 18]. Available from: https://janusgraph.org/
15. Bader A, Kopp O, Falkenthal M. Survey and Comparison of Open Source Time Series Databases. 2017.
16. Oracle. Database Data Warehousing Guide. 2017 [cited 2021 May 5]. SQL for Pattern Matching. Available from: https://docs.oracle.com/database/121/DWHSG/pattern.htm#DWHSG8956
17. Snowflake Inc. Documentation. 2021 [cited 2021 May 20]. MATCH_RECOGNIZE. Available from: https://docs.snowflake.com/en/sql-reference/constructs/match_recognize
18. Trino. Trino Documentation. 2021 [cited 2021 May 5]. MATCH_RECOGNIZE . Available from: https://trino.io/docs/current/sql/match-recognize.html
19. Camacho-Rodríguez J, Chauhan A, Gates A, Koifman E, O'Malley O, Garg V, et al. Apache Hive. In: Proceedings of the 2019 International Conference on Management of Data. New York, NY, USA: ACM; 2019. p. 1773–86.
20. doxygen. MADlib 2.1.0. 2021 [cited 2021 May 21]. User Documentation for Apache MADlib - Path. Available from: https://madlib.apache.org/docs/latest/group__grp__path.html21
21. ClickHouse. ClickHouse Documentation . 2021 [cited 2021 May 5]. Parametric Aggregate Functions. Available from: https://clickhouse.com/docs/en/sql-reference/aggregate-functions/parametric-functions
22. Vertica. Vertica Analytics Platform Version Documentation . 2021 [cited 2021 May 5]. MATCH Clause. Available from: https://www.vertica.com/docs/11.0.x/HTML/Content/Authoring/SQLReferenceManual/Statements/SELECT/MATCHClause.htm
23. teradata. Introduction to Analytics Database Analytic Functions. 2021 [cited 2021 May 31]. Database Analytic Functions. Available from: https://docs.teradata.com/r/Enterprise_IntelliFlex_VMware/Database-Analytic-Functions/Path-and-Pattern-Analysis-Functions/nPath
24. Flink. Application Development. 2022 [cited 2022 Aug 8]. Pattern Recognition. Available from: https://nightlies.apache.org/flink/flink-docs-stable/docs/dev/table/sql/queries/match_recognize/
25. Calcite. MATCH_RECOGNIZE - Calcite Documentation [Internet]. 2021 [cited 2021 May 21]. Available from: https://calcite.apache.org/docs/reference.html#match_recognize
26. Zeng M. Pattern Matching with Beam SQL [Internet]. 2020 [cited 2021 May 21]. Available from: https://beam.apache.org/blog/pattern-match-beam-sql/
27. Health Level Seven International. Health Level Seven Arden Syntax for Medical Logic Systems, Edition 3.0 [Internet]. 2023 [cited 2025 Jan 8]. Available from: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=639

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

28. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124–30.
29. Observational Health Data Sciences and Informatics. Observational Health Data Sciences and Informatics [Internet]. 2022 [cited 2025 Jan 8]. Available from: https://www.ohdsi.org/
30. openEHR. Archetype Query Language (AQL) [Internet]. 2021 [cited 2025 Jan 8]. Available from: https://specifications.openehr.org/releases/QUERY/latest/AQL.html
31. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S, McGale P, Correa C, Taylor C, Arriagada R, et al. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet. 2011 Nov 12;378(9804):1707–16.
32. Clarke M, Collins R, Darby S, Davies C, Elphinstone P, Evans V, et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005 Dec 17;366(9503):2087–106.
33. Ardizzoni A, Boni L, Tiseo M, Fossella F V, Schiller JH, Paesmans M, et al. Cisplatin-versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis. J Natl Cancer Inst. 2007 Jun 6;99(11):847–57.
34. Laxman S, Sastry PS. A survey of temporal data mining. Vol. 31, S¯ adhan¯ a. 2006.
35. Xing Z, Pei J, Keogh E. A Brief Survey on Sequence Classification. SIGKDD Explor. Newsl. 2010, ACM New York, NY, USA. Available from: https://dl.acm.org/doi/10.1145/1882471.1882478
36. Fournier-Viger P, Chun J, Lin W, Kiran RU, Koh YS, Thomas R. A Survey of Sequential Pattern Mining. Vol. 1. 2017.
37. Xu R, Wunsch D. Survey of clustering algorithms. Vol. 16, IEEE Transactions on Neural Networks. 2005. p. 645–78.
38. Warren Liao T. Clustering of time series data—a survey. Pattern Recognit. 2005 Nov;38(11):1857–74.
39. Sakoe H, Chiba S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Trans Acoust. 1978 Feb;(26):43–9.
40. Rani S, Sikka G. Recent Techniques of Clustering of Time Series Data: A Survey. Int J Comput Appl. 2012 Aug 30;52(15):1–9.
41. Vogt V, Scholz SM, Sundmacher L. Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. Eur J Public Health. 2018 Apr 1;28(2):214–9.
42. Herla F, Horton S, Mair P, Haegeli P. Snow profile alignment and similarity assessment for aggregating, clustering, and evaluating snowpack model output for avalanche forecasting. Geosci Model Dev. 2021 Jan 15;14(1):239–58.
43. Moreau C, Chanson A, Peralta V, Devogele T, Moreau C, Peralta V, et al. Clustering Sequences of Multi-dimensional Sets of Semantic Elements. 2021;384–91. Available from: https://doi.org/10.1145/3412841.3441920

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

44. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018 Feb 9; Available from: http://arxiv.org/abs/1802.03426
45. Herrmann M, Kazempour D, Scheipl F, Kröger P. Enhancing cluster analysis via topological manifold learning. Data Min Knowl Discov. 2024 May 29;38(3):840–87.
46. Allaoui M, Kherfi ML, Cheriet A. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In 2020. p. 317–25.
47. 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 Aug 9;20(Doc11).
48. Brandt PS, Kiefer RC, Pacheco JA, Adekkanattu P, Sholle ET, Ahmad FS, et al. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. Learn Health Syst. 2020 Oct 25;4(4).
49. U.S. Centers for Medicare & Medicaid Services. CMS [Internet]. 2024 [cited 2024 Mar 14]. Available from: https://www.cms.gov/
50. National Committee for Quality Assurance. HEDIS and Performance Measurement [Internet]. 2024 [cited 2024 Mar 14]. Available from: https://www.ncqa.org/hedis/

V Anlagen

Anlage 1:	Evaluationsbericht PMAW
Anlage 2:	Evaluationsbericht Data Mining
Anlage 3:	PMAW CQL und SQL Vergleichs Anfragen der Evaluation
Anlage 4:	Anforderungskatalog PMAW
Anlage 5:	Konzept Analysemodell und Abfragewerkzeug PMAW
Anlage 6:	Konzept Data Mining Werkzeug

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

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

Gefördert vom G-BA Innovationsfond
FKZ: 01VSF20018

EVALUATIONSBERICHT PMAW

11. MÄRZ 2025

Version v1.1

FLORIAN OESTERLING, STEFANIE SCHULZE
LKR NRW
Gesundheitscampus 10, 44801 Bochum

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

Inhalt

1. Hintergrund und Fragestellung

Krebsregister spielen eine zentrale Rolle in der onkologischen Versorgung und Forschung in Deutschland. Sie dienen der Erhebung, Analyse und Auswertung von umfassende Daten zu Krebserkrankungen, darunter auch Behandlungsverläufe, um wissenschaftliche Erkenntnisse für die medizinische Forschung und Versorgung zu gewinnen.

Seit dem Krebsregistergesetz von 1994 wurden in ganz Deutschland epidemiologische Krebsregister eingerichtet. Ihre Hauptaufgabe besteht in der Erfassung systematischer Daten über das Auftreten (Inzidenz) und die Verbreitung (Prävalenz) von Krebserkrankungen und deren spezifischen Merkmalen in der Bevölkerung. Auch die Erfassung von Sterblichkeit (Mortalität) und die Analyse der erhobenen Daten zur Erforschung von Trends von Krebserkrankungen in der Bevölkerung gehört dazu.

Erst seit 2013 besteht eine bundesweite Verpflichtung zur Einrichtung klinischer Krebsregister in den einzelnen Bundesländern. Die Aufgaben und Pflichten der klinischen Krebsregister sind im Krebsfrüherkennungs- und -registergesetz (KFRG, § 65c SGB V) sowie in den jeweiligen Landeskrebsregistergesetzen geregelt.
Darunter fallen auch Qualitätssicherung und Unterstützung der onkologischen Forschung, sowie eine Berichtspflicht gegenüber Vertretern der Gesundheitsplanung und Politikberatung.

Mit der Etablierung von klinischen Krebsregistern in Deutschland und den im Rahmen der klinischen Krebsregistrierung erhobenen Informationen ändern sich auch die Anforderungen an Analysetools, die eingesetzt werden, um die neuartigen Daten auswertbar zu machen. Während sich in den schon deutlich länger bestehenden epidemiologischen Krebsregistern jeder Datensatz auf die beste Information zu einem Tumor zum Zeitpunkt der Diagnose beschränkte, beinhalten die klinischen Daten nun auch Informationen zu Verlaufsereignissen wie Therapien und Verlaufsuntersuchungen. Durch diese zusätzliche zeitliche Komponente gewin-

nen die Daten an Komplexität, allerdings ermöglicht diese auch die Beantwortung von innovativer Forschungsfragen, nämlich solchen, die die zeitlichen Zusammenhänge von Krebserkrankungen und Therapien untersuchen.

Zu den Aufgaben der klinischen Krebsregister gehört die Qualitätssicherung der onkologischen Forschung. Durch den Abgleich von Behandlungsdaten mit den aktuellen S3-Leitlinien können Abweichungen identifiziert und Handlungsempfehlungen abgeleitet werden. Zudem ermöglichen Registerauswertungen Vergleiche zwischen verschiedenen Einrichtungen und Regionen.

S3-Leitlinien sind evidenzbasierte medizinische Leitlinien, die sowohl auf einer umfassenden wissenschaftlichen Analyse als auch auf Expertenkonsens basieren. Sie werden von Fachgesellschaften wie der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) in Zusammenarbeit mit der Deutschen Krebsgesellschaft (DKG) und anderen Organisationen entwickelt. Sie enthalten systematisch erarbeitete Empfehlungen für Diagnostik, Therapie und Nachsorge von Krebserkrankungen. Die S3-Klassifikation bedeutet, dass diese Leitlinien sowohl auf einer umfassenden wissenschaftlichen Analyse als auch auf Expertenkonsens basieren. Sie werden von Fachgesellschaften wie der Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) in Zusammenarbeit mit der Deutschen Krebsgesellschaft (DKG) und anderen Organisationen entwickelt.

Qualitätsindikatoren (QIs) in der Onkologie sind messbare Kriterien zur Beurteilung und Sicherstellung der Behandlungsqualität bei Krebserkrankungen, die auf den Empfehlungen der S3-Leitlinien basieren und eine objektive Bewertung der onkologischen Versorgung ermöglichen. Dazu gehört beispielsweise die Rate der Patienten, welche eine leitliniengerechte Therapie erhalten. Diese Indikatoren dienen als Grundlage für die Qualitätssicherung in zertifizierten onkologischen Zentren und werden regelmäßig von Krebsregistern und Fachgesellschaften überprüft. Die Krebsregister geben durch QIs auch eine Rückmeldung zur Meldequalität. Wenn beispielsweise eine bestimmte in den Leitlinien vorgesehene Therapie nicht ordnungsgemäß gemeldet wird, so wird der Meldestelle eine schlechte Rate für den entsprechenden QI zurückgespielt. Diese Verwendung der QIs vom Krebsregister hat keine direkte Auswirkung auf die Behandlung der Patienten. Auch wenn bei korrekter Meldung bei einem Melder schlechte QI-Erfüllungsquoten auftreten, ist es nicht die Aufgabe der Krebsregister konkrete Behandlungsempfehlungen auszusprechen, sondern nur die Information über die Qualität der im Krebsregister vorliegenden Daten.

Eine weitere gesetzlich verankerte Aufgabe besteht in der Unterstützung der onkologischen Forschung. Krebsregister stellen für Wissenschaftler pseudonymisierte Registerdaten oder aggregierte Analysen von Basisdaten und Therapieverläufen zusammen, die daraus neue Erkenntnisse zur Entstehung, Verbreitung und Behandlung von Krebserkrankungen gewinnen können (Externe Datennutzung).

Der Berichtspflicht gegenüber Vertretern der Gesundheitsplanung und Politikberatung wird in den Krebsregistern auf verschiedene Weisen nachgekommen, beispielsweise durch Jahresberichte im pdf-Format oder interaktive Online-Tools, die oft auf einem Datawarehouse basieren. Ein Data Warehouse ist ein zentrales, strukturiertes System zur Speicherung, Integration und Analyse großer Datenmengen aus verschiedenen Quellen. Daten aus dem Krebsregister

und möglichen anderen Quellen wie z. B. Meldeämtern werden in das Data Warehouse geladen, wo sie bereinigt, vereinheitlicht und für analytische Zwecke optimiert werden. Mithilfe sogenannter ETL-Prozesse (Extract, Transform, Load) werden die Daten extrahiert, in ein einheitliches Format transformiert und schließlich in das zentrale System geladen. Dort können sie mit speziellen Analyse- und Berichtstools ausgewertet werden. Im Fall eines Online-Krebsjahresberichts werden beispielsweise Informationen zu Neuerkrankungen, Überlebensraten, Therapieverläufen oder regionalen Unterschieden in der Versorgung aus dem Data Warehouse abgerufen. Klinische und epidemiologische Daten werden dazu eng verknüpft und zusammen aufbereitet. Diese Daten fließen in interaktive Dashboards oder Berichte ein, die online verfügbar sind und von Fachleuten, Forschern oder der Öffentlichkeit genutzt werden können (Routineauswertungen).

Das im ersten Abschnitt des SePaMiM Projekts entwickelte Auswertungswerkzeug soll im Krebsregister die Einteilung von Patienten von zeitlichen Abfolgen von Ereignissen in der Patientenhistorie erleichtern. Standardmäßig werden dazu Code-basierte Abfragesprachen wie die Structured Query Language (SQL) direkt auf einer Datenbank oder andere Programmier- oder Datenanalysesprachen (z.B. R, SAS, SPSS, …) auf einem Datenbankexport eingesetzt. Diese Verfahren sind aufwendig und die Anwendung erfordert sowohl sehr gute Kenntnis der gewählten Methode als auch umfangreiches Verständnis für die Datenstrukturen. Daher können die nur von Experten im Bereich Data Science oder Datenbankmanagement durchgeführt werden. Im Krebsregistern arbeiten außer diesen Personen aber noch viele weitere, die für ihre Aufgaben im Bereich der Qualitätssicherung und Melder-Betreuung des Öfteren Zugriff auf die Daten brauchen. Dieser Personengruppe soll durch den Einsatz von Pattern-Matching Verfahren ermöglicht werden, selbstständig Datenabfragen zu tätigen ohne fortgeschrittene Programmierkenntnisse. Ähnlich wie mit SQL soll auch bei der Arbeit mit diesem Tool ein großes Ausmaß an Flexibilität erhalten bleiben, um nicht nur eine Fragestellung beantworten zu können, sondern gemäß der Vielfalt der Aufgaben im Krebsregister viele unterschiedliche. Dazu wurde von OFFIS das Pattern Matching Auswertungswerkzeug (PMAW), basierend auf der Clinical Quality Language (CQL), entwickelt.

Die Grundvoraussetzung für den Erfolg des PMAW ist die Korrektheit des Ergebnisses, beziehungsweise die Übereinstimmung der durch das PMAW erzeugten Daten mit den Daten, die durch das Standardwerkzeug im Krebsregister generiert werden. Daher wird im vorliegenden Bericht der Erfolg des PMAW quantifiziert. Dazu wird das AW mit herkömmlichen und in klinischen Krebsregistern bereits gelebten Auswertungsmethoden verglichen. Zum Zwecke dieses Vergleichs werden zwei klinisch relevante Fallbeispiele definiert, ein drittes Fallbeispiel, das eine Prüfung der Datenqualität zum Inhalt hat, sowie zwei weitere aus dem Alltagsgeschäft von Krebsregistern. Beim vierten Beispiel handelt sich um eine typische Datenanfrage eines externen Wissenschaftlers und beim fünften geht es um die Bereitstellung von aggregierten Daten für einen Online-Jahresbericht. All diese Beispiele sind Standardaufgaben im Krebsregister, für die es noch viele weitere Beispiele gibt.

Die Evaluation der ersten drei Fallbeispiele wurde auf dem von OFFIS aus verschlüsselten pseudonymisierten klinischen Best-Of-Daten des LKRs erstellten Prototypen durchgeführt. Da sich diese bis auf die verschlüsselten Identitätsnummern und einer zufällig bestimmten Zahl,

die zu allen Datumwerten addiert wurde, nicht von den Originaldaten unterscheiden, wurde darauf verzichtet die Evaluation vor Ort auf den Originaldaten zu wiederholen.

2. Methode

2.1 Definition von Erfolgsparametern für das Pattern-Matching Auswertungswerkzeug

Aus fachlicher Sicht muss das PMAW die Möglichkeit bieten, unterschiedliche Filterkriterien auf den Gesamtdatenbestand der klinischen Krebsregisterdaten anzuwenden um

0. eine Basiskohorte zu definieren, typischerweise anhand einer interessierenden Tumorentität entsprechend eines ICD-10 Diagnosecodes oder einer Gruppe von Codes und dann
1. ausgehend von 1. Subgruppen von Patienten mit definierten klinischen Verläufen innerhalb dieser Kohorte auszuwählen.

Die Zuordnung von Patienten zu klinischen Verläufen wird anhand verschiedener Evaluationskriterien nach der Anwendung der Filterkriterien bewertet. Bei allen Kriterien wird die Fallauswahl mittels standardisierter SQL Datenbankabfragen, wie sie im LKR NRW derzeit genutzt werden, als Referenz für den Vergleich mit dem PMAW herangezogen. Grundlage für die Bewertung der Güte der Klassifikation durch das PMAW bildet die in Tabelle 1 abgebildete Wahrheitsmatrix. Für alle Abfragen wurden als Grundgesamtheit *N* alle in den jeweiligen Datenbanken der Krebsregister (NRW: N = 1.116.547, Niedersachsen: N = 521.361) enthaltenen Krebsfälle ohne weitere Einschränkungen verwendet.

Tabelle 1: Wahrheitsmatrix für die Klassifikation von Patienten in definierte klinische Verläufe. Die standardisierten SQL Abfragen, die bereits im LKR NRW für die Einteilung benutzt werden, stellen die Referenz oder wahre Klasse der Patienten dar.

Wahrheitsmatrix	**PMAW**: Fall entspricht gewünschtem Verlauf	**PMAW**: Fall entspricht nicht gewünschtem Verlauf
SQL: Fall entspricht gewünschtem Verlauf	r_p	f_n
SQL: Fall entspricht nicht gewünschtem Verlauf	f_p	r_n

Die folgenden Kriterien wurden definiert und sollen für jede Gruppeneinteilung der in Abschnitt 3 beschriebenen Fallbeispielen angewendet werden:

1. **Manuelle Prüfung** von Abweichungen werden Stichprobenartig bei Fällen durchgeführt, bei denen die Klassifikation des PMAW von der Referenz abweicht (außerhalb der Hauptdiagonale der Wahrheitsmatrix). Dabei werden Experten aus dem LKR NRW die klinischen Verläufe der betreffenden Patienten manuell nachvollziehen und ggf.

Anpassungen an den SQL Abfragen vornehmen, wenn das PMAW den Fall korrekt klassifiziert hat.

2. **Prozentuale Abweichung** der durch das Pattern-Matching-Auswertungswerkzeug (PMAW) und der mit der Referenzmethode ermittelten Fallzahl

$$\%Fehler = \frac{|N_{PMAW} - N_{Referenz}|}{N_{Referenz}} \times 100$$

3. **Precision** (PV) des Auswertungswerkzeugs

   Die Patienten, die das PMAW einem klinischen Verlauf zuordnet, können entweder richtig diesem Verlauf zugeordnet sein (richtig positiv, r_p) oder fälschlich diesem Verlauf zugeordnet sein (falsch positiv, f_p). Die Genauigkeit oder Precision wird als der Anteil der korrekt einem gewünschten klinischen Verlauf zugeordneten Patienten an der Summe aller diesem Verlauf zugeordneten Patienten definiert:

$$Precision = \frac{r_p}{r_p + f_p}$$

4. **Recall** (NV) des Auswertungswerkzeugs

   Die Sensitivität oder Recall gibt an, wie viele tatsächlich zu dem interessierenden Verlauf gehörenden Patienten ($r_p + f_n$) durch das AW korrekt zugeordnet werden konnten (r_p):

$$Recall = \frac{r_p}{r_p + f_n}$$

5. **F_1-Score** des Auswertungswerkzeugs

   Der F_1-Score ist als das harmonische Mittel von Precision und Recall definiert und vereint beide Maße in einem einzigen Score (1) .Der F-Score ist robust gegenüber unausgewogenen Klassen in einem Datensatz, wie es bei den hier relevanten Fragestellungen häufig der Fall sein wird. Dabei wird ein Großteil der Patienten nicht einem bestimmten klinischen Verlauf entsprechen und nur ein kleiner Teil der Fälle, der diesem Verlauf entspricht, muss identifiziert werden.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6. **Vergleich des Survival Outcomes** bei Gruppeneinteilung mit PMAW und Referenz

   Bei den Überlebenszeitanalysen aus den Fallbeispielen werden zunächst die Ergebnisse (Milestone Survival bei 5-Jahren Follow-up bzw. Hazard Ratio) zwischen den durch die Referenzmethode und den durch das PMAW identifizierten Gruppen verglichen, um den Einfluss der Klassifikationsmethode auf den klinisch relevanten Endpunkt zu untersuchen. Zudem wird die Analyse in den beiden Gruppen der abweichend

klassifizierten Fälle durchgeführt um zu untersuchen, ob es systematische Unterschiede in der Prognose der richtig und falsch klassifizierten Fälle gibt.

7. **Analyse von prognostischen Faktoren für Progression**

   Bei der Identifikation von prognostischen Faktoren für das Auftreten eines klinischen Ereignisses wird nicht der Endpunkt direkt (Odds Ratio für Event) sondern die Verteilung der Prognostischen Eigenschaften zwischen den richtig und falsch klassifizierten Gruppen verglichen, um zu untersuchen, ob es systematische Unterschiede zwischen den durch das PMAW identifizierten Fällen und der Referenz gibt.

2.2 Festgelegte Fallbeispiele zur Evaluation

Eine Aufgabe der klinischen Krebsregister ist es, die onkologische Versorgungsrealität anhand von Krebsregisterdaten abzubilden, die gesammelten Daten auszuwerten und Ergebnisse an die meldenden Einrichtungen zurückzumelden, um so zur Verbesserung der Versorgung beizutragen. Beispiele für Auswertungen, die diesen Zielen dienen sind das Überprüfen einer leitliniengerechten Therapie sowie das Identifizieren von Patientengruppen, die besonders von bestimmten Therapieprotokollen profitieren oder ein erhöhtes Risiko für unerwünschte Ereignisse wie Therapiekomplikationen oder Rezidive haben. Das hier untersuchte Auswertungswerkzeug soll bei der Auswertung dieser klinischen Fragestellungen helfen, indem es eine nutzerfreundliche und korrekte Einteilung eines Patientenkollektivs in verschiedene klinische Verläufe erlaubt. Anhand von zwei klinisch orientierten Fragestellungen wird im folgenden Abschnitt eine solche Einteilung vorgenommen und die Ergebnisse anhand der zuvor definierten Evaluationsparameter gemessen. Eine aussagekräftige Analyse der klinischen Fragestellungen kann nur gelingen, wenn die zugrundeliegenden Daten eine hohe Qualität aufweisen. Bei dem dritten Auswertungsbeispiel steht daher nicht eine klinische Fragestellung, sondern die Prüfung der Datenqualität im Vordergrund.

Eine weitere Aufgabe der Krebsregister ist die Bereitstellung von ausgewählten Daten für externe Nutzer, insbesondere im Bereich der öffentlichen Gesundheit und Wissenschaft. Dabei lassen sich vier Arten der Datennutzung unterscheiden: individueller Patientenzugriff, Kohortenabgleich, aggregierte Daten und anonymisierte Einzelfalldaten. Individuelle Anfragen und Kohortenabgleiche basieren auf Merkmalen wie ICD-10-Diagnosecode, Alter oder Geschlecht, während aggregierte Maße wie Raten oder Fallzahlen häufig in Routinedokumenten und Berichten veröffentlicht werden. Für Forschungszwecke können anonymisierte Einzelfalldaten unter Angabe klarer Kriterien bereitgestellt werden. Um die Datensicherheit zu gewährleisten, werden nur die notwendigen Attribute mit angemessener Granularität exportiert. Die endgültige Festlegung der Kriterien und Attribute erfolgt in Absprache mit den Antragstellern.

Für die Evaluation wurden zwei Anfragen zur externen Datennutzung ausgewählt. Beide Anfragen erfolgten im Rahmen einer Studie, die den Einfluss des Zeitintervalls zwischen Diagnose und Beginn der tumorspezifischen Therapie auf das Behandlungsergebnis bei Patienten mit fortgeschrittenen Krebsdiagnosen untersuchte. Die betrachteten Krebsdiagnosen waren kolorektaler Krebs und Brustkrebs.

Das Berichtswesen der deutschen Krebsregister informiert über Krebsfälle in ihren Einzugsgebieten, beispielsweise durch Jahresberichte oder interaktive Online-Tools. Der Online-Bericht bietet einen Überblick über die Krebssituation in Niedersachsen anhand von Merkmalen wie Alter, Geschlecht, Wohnort, Tumorart und Therapie. Analysen wurden mit CQL für zwei Berichtsseiten umgesetzt: eine zeigt neue Krebsfälle nach Diagnosegruppe, Geschlecht und Wohnort, die andere liefert Daten zu Operationen, ergänzt um UICC-Stadium und behandlungsspezifische Merkmale.

Diese letzten drei Fallbeispiele waren im ursprünglichen Evaluationskonzept nicht vorgesehen, da sie erst nach dem Erstellen des Konzepts im Rahmen einer Publikation mit den Daten der klinischen Landesauswertungsstelle (KLast) Niedersachsen eingeschlossen wurden.

3. Ergebnisse

Im Folgenden wird auf die Ergebnisse der Fallbeispiele eingegangen. Dabei werden auch die verwendeten Filter kurz erläutert. Die für die Berechnung verwendeten Skripte befinden sich in der Softwaresammlung.

3.1. Qualitätsindikatoren

Fallbeispiel 1 a: Mammakarzinom QI 8 - Durchgeführte Strahlentherapie nach BET

(basierend auf dem Datensatz des LKR-NRW, N = 1.160.547 Fälle)

Dieser QI stammt aus der S3- Leitlinie Mammakarzinom Version 4.4, Juni 2021. Nach brusterhaltender Operation wegen eines invasiven Karzinoms soll eine Bestrahlung der betroffenen Brust durchgeführt werden. Das Qualitätsziel ist eine adäquate Rate an Bestrahlungen nach BET bei Patienten mit Ersterkrankung invasives Mammakarzinom.

Änderungen zum Evaluationskonzept (04/2022)

Da die Implementierung der auf den S3-Leitlinien zur Behandlung des Mammakarzinoms beruhenden QIs wie die aller weiteren QIs beständig weiterentwickelt wird haben wir uns dazu entschlossen, auch bei der Evaluation den neusten Stand des QIs zu verwenden.

Die verwendeten Filterkriterien sehen nun wie folgt aus:

Nenner: Patienten mit Primärerkrankung invasives Mammakarzinom und BET

- ICD-10 Diagnosecode(s): C50.0 – C50.9
- Diagnosedatum ≥ 01.04.2016
- Histologie zu C50*: 8010/3-8570/3
- Bei Diagnose nicht metastasiert (TNM M ≠ 1)
- Mindestens eine OP mit OPS Code 5-Steller = 5-870 vorhanden (Definition BET)
- Keine OP mit OPS Codes für Mastektomie (5-872, 5-874, 5-877)

Zähler: alle Kriterien des Nenners UND erfolgte Bestrahlung der Brust

- Mindestens eine Meldung zu Bestrahlung, Zielgebiet: 3.1 (Mamma als Ganzbrust) oder 3.2 (Mamma als Teilbrust)
- Stellung der Bestrahlung zur BET = adjuvant ODER Beginndatum der Bestrahlung ≥ OP-Datum UND Beginndatum ≤ OP-Datum + 8 Monate

Ergebnisse

Tabelle 2: Ergebnisse mittels der SQL-Abfrage und des Pattern-Matching-Auswertungswerkzeugs (PMAW), Fallbeispiel 1a.

	SQL	PMAW
Therapieverlauf 1 (Nenner)	41.134	41.134
Therapieverlauf 2 (Zähler)	19.078	19.078

Tabelle 3: Wahrheitsmatrix für Fallbeispiel 1a – Nenner.

Wahrheitsmatrix	**PMAW**: Fall entspricht gewünschtem Verlauf	**PMAW**: Fall entspricht nicht gewünschtem Verlauf
SQL: Fall entspricht gewünschtem Verlauf	r_p =41.134	f_n= 0
SQL: Fall entspricht nicht gewünschtem Verlauf	f_p= 0	r_n=1.119.413

Tabelle 4: Wahrheitsmatrix für Fallbeispiel 1a – Zähler.

Wahrheitsmatrix	**PMAW**: Fall entspricht gewünschtem Verlauf	**PMAW**: Fall entspricht nicht gewünschtem Verlauf
SQL: Fall entspricht gewünschtem Verlauf	r_p =19.078	f_n= 0
SQL: Fall entspricht nicht gewünschtem Verlauf	f_p= 0	r_n=1.141.469

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** ist nicht nötig, da die Ergebnisse zu 100% übereinstimmen.

2. **Prozentuale Abweichung**:
   Nenner:

$$\%Fehler = \frac{41.134 - 41.134}{41.134} \times 100 = 0\%$$

   Zähler:

$$\%Fehler = \frac{19.078 - 19.078}{19.078} \times 100 = 0\%$$

3. **Precision** (PV) des Auswertungswerkzeugs
   Nenner:

$$Precision = \frac{41.134}{41.134 + 0} = 1$$

   Zähler:

$$Precision = \frac{19.078}{19.078 + 0} = 1$$

4. **Recall** (NV) des Auswertungswerkzeugs
   Nenner:

$$Recall = \frac{41.134}{41.134 + 0} = 1$$

   Zähler:

$$Recall = \frac{19.078}{19.078 + 0} = 1$$

5. **F_1-Score** des Auswertungswerkzeugs
   Nenner:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

   Zähler:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

6. **Vergleich des Survival Outcomes** bei Gruppeneinteilung mit PMAW und Referenz

   Da hier keine Unterschiede zwischen den Methoden festgestellt werden konnten, wurde darauf verzichtet eine Überlebenszeitanalyse durchzuführen.

7. **Analyse von prognostischen Faktoren für Progression**

   Da hier keine Unterschiede zwischen den Methoden festgestellt werden konnten, wurde darauf verzichtet nach systematischen Unterschieden zwischen den durch das PMAW identifizierten Fällen und der Referenz zu suchen.

Fallbeispiel 1b: Lungenkarzinom QI 6 - Adjuvante Cisplatin-basierte Chemotherapie bei NSCLC Stadium II

(basierend auf dem Datensatz des LKR-NRW, N = 1.160.547 Fälle)

Dieser QI stammt aus der S3- Leitlinie Lungenkarzinom Version 2.2, Juli 2023. Die Chemotherapie soll als eine Cisplatin- haltige Kombination über 4 Zyklen erfolgen. Das Qualitätsziel ist eine möglichst häufige Behandlung mit adjuvanter Cisplatin-basierter Chemotherapie bei NSCLC Stadium II b.

Änderungen zum Evaluationskonzept (04/2022)

Aus den oben genannten Gründen haben wir uns dazu entschlossen, auch hier bei der Evaluation den neusten Stand des QIs zu verwenden.

Die verwendeten Filterkriterien sehen nun wie folgt aus:

Nenner:

- ICD-10 Diagnosecode(s): C34.0 – C34.9
- Diagnosedatum ≥ 01.04.2016
- UICC = IIA, IIB
- ICD-O-3 Morphologiecode aus Gruppe NSCLC
- GeneralPerformanceState = 0, 1 (ECOG) ODER GeneralPerformanceState >70 (Karnofsky)
- Mindestens eine tumorspezifische OP zum Tumor mit
    - R-Status = R0
    - OPS Code '5-323.43', '5-323.53', '5-323.63', '5-323.73, '5-323.x3', '5-324.3', '5-324.7', '5-324.9', '5-324.b', '5-325', '5-327.1', '5-327.3', '5-327.5', '5-327.7', '5-328', '5-404', '5-406' oder '5-407' (Lymphknotendissektion)

Zähler: alle Kriterien des Nenners UND

- Systemische Therapie mit
    - Stellung zur OP = adjuvant ODER Startdatum ≤ OP-Datum + 60 Tage
    - Therapieart = CH
    - Substanz = „Cisplatin“

Zusätzlich zu diesen im QI festgelegten Kriterien haben wir hier nach einem weiteren Ereignis im Verlauf gesucht, nämlich nach dem Auftreten eines Rezidivs. Dadurch erhöht sich die Komplexität der Abfragen noch einmal.

Therapieverlauf 1: Leitliniengerechte Therapie mit Rezidiv = Zähler UND

- Verlaufsuntersuchung mit Rezidiv oder Fernmetastase
- Datum ≥ OP-Datum

Therapieverlauf 2: Leitliniengerechte Therapie ohne Rezidiv = Zähler

Ergebnisse

Tabelle 5: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 1b, QI.

	SQL	PMAW
Therapieverlauf 1 (Nenner)	944	944
Therapieverlauf 2 (Zähler)	208	208

Tabelle 6: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 1b, Rezidiv.

	SQL	PMAW
Therapieverlauf 1 (mit Rezidiv)	41	41
Therapieverlauf 2 (ohne Rezidiv)	167	167

Tabelle 7: Wahrheitsmatrix für Fallbeispiel 1b, QI – Nenner.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 1	**PMAW**: Fall entspricht Therapieverlauf 2
SQL: Fall entspricht Therapieverlauf 1	r_p =944	f_n= 0
SQL: Fall entspricht Therapieverlauf 2	f_p= 0	r_n=1.159.603

Tabelle 8: Wahrheitsmatrix für Fallbeispiel 1b, QI – Zähler.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 1	**PMAW**: Fall entspricht Therapieverlauf 2
SQL: Fall entspricht Therapieverlauf 1	r_p =208	f_n= 0
SQL: Fall entspricht Therapieverlauf 2	f_p= 0	r_n=1.160.339

Tabelle 9: Wahrheitsmatrix für Fallbeispiel 1b, Rezidiv – Rezidiv vorhanden.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 1	**PMAW**: Fall entspricht Therapieverlauf 2
SQL: Fall entspricht Therapieverlauf 1	r_p =41	f_n= 0
SQL: Fall entspricht Therapieverlauf 2	f_p= 0	r_n= 1.160.506

Tabelle 10: Wahrheitsmatrix für Fallbeispiel 1b, Rezidiv– Kein Rezidiv vorhanden.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 1	**PMAW**: Fall entspricht Therapieverlauf 2
SQL: Fall entspricht Therapieverlauf 1	r_p =167	f_n= 0
SQL: Fall entspricht Therapieverlauf 2	f_p= 0	r_n= 1.160.380

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** ist nicht nötig, da die Ergebnisse zu 100% übereinstimmen.
2. **Prozentuale Abweichung** der durch das Pattern-Matching-Auswertungswerkzeug (PMAW) und der mit der Referenzmethode ermittelten Fallzahl beträgt:

   QI
   Nenner:

$$\%Fehler = \frac{944 - 944}{944} \times 100 = 0\%$$

   Zähler:

$$\%Fehler = \frac{208 - 208}{208} \times 100 = 0\%$$

   Rezidiv

   Vorhanden:

$$\%Fehler = \frac{41 - 41}{41} \times 100 = 0\%$$

   Nicht vorhanden:

$$\%Fehler = \frac{167 - 167}{167} \times 100 = 0\%$$

3. **Precision** (PV) des Auswertungswerkzeugs

   QI

   Nenner:

$$Precision = \frac{944}{944 + 0} = 1$$

   Zähler

$$Precision = \frac{208}{208 + 0} = 1$$

Rezidiv

Vorhanden:

$$Precision = \frac{41}{41 + 0} = 1$$

Nicht vorhanden:

$$Precision = \frac{167}{167 + 0} = 1$$

4. **Recall** (NV) des Auswertungswerkzeugs

QI
Nenner:

$$Recall = \frac{944}{944 + 0} = 1$$

Zähler:

$$Recall = \frac{208}{208 + 0} = 1$$

Rezidiv
Vorhanden:

$$Recall = \frac{41}{944 + 0} = 1$$

Nicht vorhanden:

$$Recall = \frac{167}{167 + 0} = 1$$

5. **F_1-Score** des Auswertungswerkzeugs

QI
Nenner:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

Zähler:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

Rezidiv
Vorhanden:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

Nicht vorhanden:

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

6. **Vergleich des Survival Outcomes** bei Gruppeneinteilung mit PMAW und Referenz

   Da hier keine Unterschiede zwischen den Methoden festgestellt werden konnten, wurde darauf verzichtet eine Überlebenszeitanalyse durchzuführen.

7. **Analyse von prognostischen Faktoren für Progression**

   Da hier keine Unterschiede zwischen den Methoden festgestellt werden konnten, wurde darauf verzichtet nach systematischen Unterschieden zwischen den durch das PMAW identifizierten Fällen und der Referenz zu suchen.

3.2. Plausibilitätsprüfungen

(basierend auf dem Datensatz des LKR-NRW, N = 1.160.547 Fälle)

Fallbeispiel 2a: Plausibilitätsprüfung - Klinische Ereignisse nach dem Tod des Patienten

Das Todesdatum kann aus mehreren Quellen stammen, es kann direkt von einer Einrichtung gemeldet werden oder durch regelmäßige Abgleiche mit den Einwohnermeldeämtern in unsere Datenbank aufgenommen werden. Sollten mehrere Therapieereignisse nach dem Tod des Patienten von verschiedenen Meldern gefunden werden, ist zu vermuten, dass ein Datum in der Datenbank falsch ist. Neben der Möglichkeit einer falschen Meldung könnte es auch ein Hinweis auf einen Fehler bei der Zuordnung von Todesmeldungen mit Krebsregisterfällen sein.

Operationalisierung der Filterkriterien

- Todesdatum < Op-Datum
- Todesdatum < Strahlentherapie Startdatum ODER
  Todesdatum < Strahlentherapie Enddatum ODER
  Todesdatum < Bestrahlung Startdatum ODER
  Todesdatum < Bestrahlung Enddatum ODER
- Todesdatum < Systemische Therapie Startdatum ODER
  Todesdatum < Systemische Therapie Enddatum ODER

Hinweis: Das Enddatum von Therapien wird oft nicht gemeldet, daher muss auch das Startdatum in die Abfrage mit ein bezogen werden. Zu jeder Strahlentherapie werden auch einzelne Bestrahlungsdaten gelmeldet.

Es wurden keine Änderungen am Evaluationskonzept vorgenommen.

Ergebnisse

Tabelle 11: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 2a

	SQL	PMAW
Therapieverlauf 1 (OP)	176	176
Therapieverlauf 2 (Radiotherapie)	197	197
Therapieverlauf 3 (Systemtherapie)	1611	1611

Tabelle 12: Wahrheitsmatrix für Fallbeispiel 2a, Plausibilität – OP.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 1	**PMAW**: Fall entspricht nichtTherapieverlauf 1
SQL: Fall entspricht Therapieverlauf 1	r_p = 176	f_n= 0
SQL: Fall entspricht nicht Therapieverlauf 1	f_p= 0	r_n= 1.160.371

Tabelle13: Wahrheitsmatrix für Fallbeispiel 2a, Plausibilität – Radiotherapie.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 2	**PMAW**: Fall entspricht nicht Therapieverlauf 2
SQL: Fall entspricht Therapieverlauf 2	r_p = 197	f_n= 0
SQL: Fall entspricht nicht Therapieverlauf 2	f_p= 0	r_n= 1.160.350

Tabelle 14: Wahrheitsmatrix für Fallbeispiel 2a, Plausibilität – Systemtherapie.

Wahrheitsmatrix	**PMAW**: Fall entspricht Therapieverlauf 3	**PMAW**: Fall entspricht nicht Therapieverlauf 3
SQL: Fall entspricht Therapieverlauf 3	r_p = 1.611	f_n= 0
SQL: Fall entspricht nicht Therapieverlauf 3	f_p= 0	r_n= 1.158.936

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** ist nicht nötig, da die Ergebnisse zu 100% übereinstimmen.

2. **Prozentuale Abweichung**:
   OP

$$\%Fehler = \frac{176 - 176}{176} \times 100 = 0\%$$

Radiotherapie

$$\%Fehler = \frac{197 - 197}{197} \times 100 = 0\%$$

Systemtherapie

$$\%Fehler = \frac{1611 - 1611}{1611} \times 100 = 0\%$$

3. **Precision** (PV) des Auswertungswerkzeugs

   OP

$$Precision = \frac{176}{176 + 0} = 1$$

   Radiotherapie

$$Precision = \frac{197}{197 + 0} = 1$$

   Systemtherapie

$$Precision = \frac{1611}{1611 + 0} = 1$$

4. **Recall** (NV) des Auswertungswerkzeugs

   OP

$$Recall = \frac{176}{176 + 0} = 1$$

   Radiotherapie

$$Recall = \frac{197}{197 + 0} = 1$$

   Systemtherapie

$$Recall = \frac{1611}{1611 + 0} = 1$$

5. **F_1-Score** des Auswertungswerkzeugs

   OP, Radiotherapie, Systemtherapie

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

Fallbeispiel 2b: Plausibilitätsprüfung - Wechsel von palliativer zu kurativer Therapieintention

Ein Wechsel von palliativer zu kurativer Therapieintention ist unwahrscheinlich, insbesondere, wenn ein mehrfacher Wechsel zwischen diesen Intentionen gemeldet wird.

Es wurden keine Änderungen am Evaluationskonzept vorgenommen.

Ergebnisse

Tabelle 15 Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 2b.

	SQL	PMAW
Fälle mit Wechsel von palliativer zu kurativer Therapieintention	6787	6788

Tabelle 16: Wahrheitsmatrix für Fallbeispiel 2b.

Wahrheitsmatrix	**PMAW**: Fall entspricht Suchkriterien	**PMAW**: Fall entspricht Suchkriterien nicht
SQL: Fall entspricht Suchkriterien	r_p = 6786	f_n= 1
SQL: Fall entspricht Suchkriterien nicht	f_p= 2	r_n= 1.153.759

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** hat ergeben, dass diese Abweichungen darauf zurückzuführen sind, dass bei der Erstellung des CQL-Datensatzes die Start- und Enddaten von Behandlungsevents vertauscht wurden, wenn das Startdatum nach dem Enddatum lag. Dadurch entstanden die zwingend notwendigen gültigen Zeitintervalle, aber die chronologische Reihenfolge einiger Behandlungsevents wurde geändert.

2. **Prozentuale Abweichung**:

$$\%Fehler = \frac{6788 - 6787}{6787} \times 100 = 0{,}02\%$$

3. **Precision** (PV) des Auswertungswerkzeugs:

$$Precision = \frac{6786}{6786 + 2} = 0{,}999705362$$

4. **Recall** (NV) des Auswertungswerkzeugs:

$$Recall = \frac{6786}{6786 + 1} = 0{,}999852659$$

5. **F_1-Score** des Auswertungswerkzeugs:

$$F_1 = 2 \times \frac{0{,}999705362 \times 0{,}999852659}{0{,}999705362 + 0{,}999852659} = 0{,}99979006$$

Fallbeispiel 2c: Plausibilitätsprüfung - Rückgang der Tumorgröße ohne Therapie

Ein Rückgang der Tumorgröße bei zwei aufeinanderfolgenden TNM Ereignissen sollte nur vorkommen, wenn eine Therapie zwischen den Ereignissen lag, die zu dieser Reduktion geführt hat. Diese Plausibilitätsprüfung untersucht die Zahl von Patienten, bei denen es einen implausiblen Rückgang der Tumorgröße gab.

Es wurden keine Änderungen am Evaluationskonzept vorgenommen.

Ergebnisse

Tabelle 17: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 2c.

	SQL	PMAW
Fälle mit Rückgang der Tumorgröße ohne Therapie	8568	2370

Tabelle 18: Wahrheitsmatrix für die Fallbeispiel 2c.

Wahrheitsmatrix	**PMAW**: Fall entspricht Suchkriterien	**PMAW**: Fall entspricht Suchkriterien nicht
SQL: Fall entspricht Suchkriterien	r_p =1520	f_n= 7048
SQL: Fall entspricht Suchkriterien nicht	f_p=850	r_n= 1.158.177

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** hat ergeben, dass die Unterschiede auf die unterschiedliche Verarbeitung der TNM-Ereignisse beruhen. Für die meisten Auswertungen und so auch hier wird, wenn nicht explizit anders verlangt, der pathologische TNM (pTNM) dem klinischen (cTNM) vorgezogen, d.h. Konkret, dass nur beim Fehlen des pathologischen der klinische TNM herangezogen wird.
   Der für CQL aufbereitete Tumorstatus übernimmt jedoch beim Fehlen eines der beiden TNM-T-Werte den Wert des letzten Status, so dass beim Fehlen des pTNM-T nicht der cTNM-T, sondern der vorherige pTNM-T herangezogen wird. Dadurch ist zu erklären, dass ein cTNM-T-Wert, der den Suchkriterien entspricht, vom SQL Skript erkannt wird, während im CQL-Prototypen dieser nicht angezeigt wird.

2. **Prozentuale Abweichung** der durch das Pattern-Matching-Auswertungswerkzeug (PMAW) und der mit der Referenzmethode ermittelten Fallzahl beträgt:

$$\%Fehler = \frac{1520 - 8568}{8568} \times 100 = -82{,}25957049\%$$

3. **Precision** (PV) des Auswertungswerkzeugs:

$$Precision = \frac{1520}{1520 + 850} = 0{,}9470404984$$

4. **Recall** (NV) des Auswertungswerkzeugs:

$$Recall = \frac{1520}{1520 + 7824} = 0{,}1626712329$$

5. **F_1-Score** des Auswertungswerkzeugs:

$$F_1 = 2 \times \frac{1 \times 0{,}232893799}{1 + 0{,}232893799} = 0{,}37780026$$

3.3. Externe Datennutzung

Fallbeispiel 3a: Datenlieferung: Brustkrebs (C50)

(basierend auf dem Datensatz der KLast Niedersachsen, N = 521.361)

Bei der Datenanfrage an die KLast durch eine Universitätsmedizin ging es um die Durchführung einer Studie mit dem Ziel den Einfluss des Therapiezeitpunkts auf das Überleben bei Patienten mit fortgeschrittenem Mammakarzinom (UICC IV) zu untersuchen.

Es wurde nach ICD-Code des Tumors und dem Vorhandensein einer palliativen systematischen Therapie, sowie einigen typischen weiteren Merkmalen gefiltert, darunter Diagnosezeitraum und Geschlecht des Patienten.

Ergebnisse

Tabelle 19: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 3a

	SQL	PMAW
Fälle für den Datenexport	1206	1206

Tabelle 20: Wahrheitsmatrix für Fallbeispiel 3a .

Wahrheitsmatrix	**PMAW**: Fall entspricht Suchkriterien	**PMAW**: Fall entspricht Suchkriterien nicht
SQL: Fall entspricht Suchkriterien	r_p =1206	f_n= 0
SQL: Fall entspricht Suchkriterien nicht	f_p=0	r_n= 520.155

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** ist nicht nötig, da die Ergebnisse zu 100% übereinstimmen.

2. **Prozentuale Abweichung**:

$$\%Fehler = \frac{1206 - 1206}{1206} \times 100 = 0\%$$

3. **Precision** (PV) des Auswertungswerkzeugs

$$Precision = \frac{1206}{1206 + 0} = 1$$

4. **Recall** (NV) des Auswertungswerkzeugs

$$Recall = \frac{1206}{1206 + 0} = 1$$

5. **F_1-Score** des Auswertungswerkzeugs

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

Fallbeispiel 3b: Datenlieferung: Darmkrebs (C18-20)

Bei der Datenanfrage an die KLast durch eine Universitätsmedizin ging es um die Durchführung einer Studie mit dem Ziel den Einfluss des Therapiezeitpunkts auf das Überleben bei Patienten mit fortgeschrittenem kolorektalen Karzinom (UICC IV) zu untersuchen.

Es wurde nach ICD-Code des Tumors und dem Vorhandensein einer palliativen systematischen Therapie, sowie einigen typischen weiteren Merkmalen gefiltert, darunter Diagnosezeitraum und Geschlecht des Patienten.

Ergebnisse

Tabelle 21: Ergebnisse mittels der SQL-Abfrage und des PMAW, Fallbeispiel 3b

	SQL	PMAW
Fälle für den Datenexport	1685	1685

Tabelle 22: Wahrheitsmatrix für Fallbeispiel 3b .

Wahrheitsmatrix	**PMAW**: Fall entspricht Suchkriterien	**PMAW**: Fall entspricht Suchkriterien nicht
SQL: Fall entspricht Suchkriterien	r_p = 1685	f_n= 0
SQL: Fall entspricht Suchkriterien nicht	f_p=0	r_n= 519.676

Die folgenden Kriterien wurden im Evaluationskonzept definiert:

1. **Manuelle Prüfung** ist nicht nötig, da die Ergebnisse zu 100% übereinstimmen.

2. **Prozentuale Abweichung**:

$$\%Fehler = \frac{1685 - 1685}{1685} \times 100 = 0\%$$

3. **Precision** (PV) des Auswertungswerkzeugs

$$Precision = \frac{1685}{1685 + 0} = 1$$

4. **Recall** (NV) des Auswertungswerkzeugs

$$Recall = \frac{1685}{1685 + 0} = 1$$

5. **F_1-Score** des Auswertungswerkzeugs

$$F_1 = 2 \times \frac{1 \times 1}{1 + 1} = 1$$

3.4. Routineauswertungen

(basierend auf dem Datensatz der KLast Niedersachsen, N = 521.361)

Der interaktive Online-Bericht der KLast bietet einen umfassenden Überblick über die Krebssituation in Niedersachsen. Für die Evaluation wurden zwei Seiten des Jahresberichts mit CQL analysiert: eine Seite zu neuen Krebsfällen (nach Diagnosegruppe, Geschlecht, Diagnosejahr und Wohnort – Fallbeispiel 4a Inzidenz) und eine Seite zu Operationen (mit zusätzlichen Merkmalen wie UICC-Stadium und Zeit bis zur ersten Operation – Fallbeispiel 4b Operation). Die

Daten wurden in CQL gefiltert, als JSON exportiert und in R weiterverarbeitet, um die erforderlichen Kennzahlen zu berechnen. Für beide Anfragen bestand die untersuchte Kohorte aus 152.590 Krebsfällen. Für beide Anfragen stimmten die Ergebnisse mit den bisherigen Methoden der KLast (Erstellung von MDX-Abfragen in einem Data Warehouse und anschließende Verarbeitung mit einem C#-Tool) exakt überein.

4. Diskussion

Wie die Zahlen der ersten beiden Fallbeispiele (1a und 1b) zeigen, ist der Prototyp in der Lage, die exakt gleichen Ergebnisse zu liefern, wie eine SQL-Abfrage. Somit sind die Anforderungen an das Tool zu 100% erfüllt.

Die Unterschiede in Fallbeispiel 2c sind dadurch bedingt, dass das CQL-Modell einen höheren Anspruch an die Datenqualität und -plausibilität hat. Die Daten mussten daher vor der Evaluation noch weiter aufbereitet werden und wurden damit auch verändert.

Die hohe Diskrepanz zwischen den Ergebnissen des SQL-Skript und des Prototyps kommt, wie im Ergebnisteil bereits beschrieben, durch die Einführung des Tumorstatus zu Stande. Da durch diese Veränderungen jedoch weniger implausible Fälle, hier: "Wunderheilungen", bei denen sich der TNM-T ohne Behandlung verkleinert und man davon ausgehen muss, dass es sich bei einem Großteil um Datenartefakte und nicht um Spontanheilungen handeln wird, gefunden werden, könnte es durchaus Sinn machen den Tumorstatus auch bei entsprechenden Auswertungen einzusetzen. Allerdings muss man beachten, dass das Fehlen des verlässlicheren pTNM-Status nur eine Quelle für die hier gefundenen Implausibilitäten ist. Ebenso kann eine nicht gemeldete Therapie zu dieser scheinbar unerwarteten Verringerung des TNM-T-Werts führen.

Alles in allem hat CQL sich als nützliches Werkzeug für die Kohortenbildung und Datenanalysen in Krebsregistern erwiesen, da es komplexe Kriterien und zeitliche Beziehungen abbilden kann. Es ermöglicht standardisierte und lesbare Abfragen, unterstützt die Formulierung von Qualitätsindikatoren und erleichtert die Datenbereitstellung für externe Anfragen. Zukünftige Verbesserungen des CQL-Editors und der Nutzerfreundlichkeit sind geplant.

Wie bereits oben erwähnt, dient dieses Dokument dazu, die technischen Anforderungen an das Tool zu evaluieren. Der praktische Nutzen im Krebsregister wurde nicht metrisch erfasst und analysiert. Im Rahmen einer internen Veranstaltung wurde das Tool vorgestellt und mit positivem Ergebnis diskutiert.
Der klinische Nutzen ergibt sich aus den Aufgaben der klinischen Krebsregistrierung, die über Qualitätsindikatoren die Versorgungsqualität an die Melder zurückspiegelt mit dem Ziel, dass diese ihre Qualität verbessern. Zudem werden Daten für externe (klinische) Forschungsanfragen und die Versorgungsforschung bereitgestellt. Damit lässt sich z.B., im Gegensatz zu klinischer Versorgung, die Effektivität von Therapien in einem Real-World-Setting vergleichen. Die Routineberichterstattung dient der Politikberatung sowie Gesundheitsplanung und ist damit eher von informierender, strategischer Bedeutung. PMAW unterstützt die Krebsregister bei allen diesen Aufgaben und erleichtert diese.

Bochum, 11.03.2025

Florian Oesterling
Verantwortlicher Methodiker
LKR NRW, Bochum

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

Gefördert vom G-BA Innovationsfond
FKZ: 01VSF20018

EVALUATIONSBERICHT DATA MINING

10. JANUAR 2025

Version v2.0

FLORIAN OESTERLING, STEFANIE SCHULZE
LKR NRW
Gesundheitscampus 10, 44801 Bochum

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

Inhalt

1. Hintergrund und Fragestellung

Mit Hilfe des in den ersten anderthalb Jahren des SePaMiM-Projekts entwickelten Pattern-Matching-Auswertungswerkzeugs (PMAW) können die aufbereiteten Daten des klinischen Krebsregisters NRW nun auf bestimmte Behandlungsverläufe wie z.B. leitliniengerechten Therapieabfolgen geprüft werden.

Ebenso interessant ist es, die komplexen und von vielen möglichen Ereignissen gekennzeichneten Behandlungsverlaufsdaten auf bisher unbekannte Muster zu durchsuchen. Daher soll nun im zweiten Teil des SePaMiM-Projekts ein Data-Mining-Werkzeug entwickelt werden, das diesen Zweck erfüllt.

2. Methoden

In enger Zusammenarbeit zwischen OFFIS und dem LKR NRW wurde entschieden, für den zweiten Projektteil zunächst ein unüberwachtes Lernverfahren, nämlich Clustering, zu verwenden. Ziel soll es sein, nach sorgfältigem Vorfiltern von Tumoreigenschaften wie ICD-Code, Morphologie und Krankheitsstadium Patientenkohorten anhand von Behandlungsverläufen in einzelnen Clustern zu identifizieren.

Die Evaluation des jeweiligen Modells findet in drei Schritten statt. Der erste Evaluationsschritt ist die Überprüfung der Modellergebnisse anhand von technischen Kriterien, die die Funktionalität des Modells überprüfen. In einem zweiten Schritt sollen verschiedene inhaltliche Prüfungen durch Fachexperten des LKR vorgenommen werden. Ziel ist hierbei die klinische Relevanz und Interpretierbarkeit der ermittelten Cluster direkt durch potentielle Anwender zu evaluieren. Zuletzt wird ein Ansatz herangezogen werden, um klinisch relevante Unterschiede zwischen den Clustern anhand von konkreten klinischen Fragestellungen zu untersuchen. Dabei haben wir uns für eine Überlebenszeitanalyse der Patienten im inter-Cluster-Vergleich entschieden.

2.1 Technische Qualität und Plausibilität

Zunächst werden einfache Qualitätschecks durchgeführt. Die zurückgelieferten Ergebnisse sollten mit den ursprünglichen Daten eindeutig abstimmbar sein Es wird außerdem geprüft, ob alle übergebenen Beobachtungen im Ergebnis vorhanden sind und einem Cluster zugeordnet wurden, oder ob es solche gibt, für die keine Zuordnung erfolgen konnte. Falls eine Zuordnung nicht erfolgen konnte, sollen die Gründe ermittelt werden. Beispielsweise können fehlende Werte in den Ursprungsdaten dazu führen, dass ein Clusteralgorithmus keine Zuordnung durchführen kann.

Anschließend werden mathematische Plausibilitätschecks durchgeführt. Es wird anhand von metrischen Werten bestimmt, wie hoch die Clusterähnlichkeit innerhalb eines Clusters und die Unterschiedlichkeit zwischen den Clustern sind. Eine hohe Variabilität der Werte innerhalb eines Clusters kann ein Indikator dafür sein, dass aus dem einen Cluster besser zwei gemacht werden sollten, während eine geringe Variabilität zwischen Clustern ein Hinweis darauf sein kann, dass zu viele Cluster gebildet wurden. Wünschenswert wären Vergleichswerte, die beispielsweise dadurch ermittelt werden könnten, dass ein Clusteralgorithmus mit mehreren Clusteranzahlen durchgeführt wird. Eine sinnvolle Ermittlung der Clusteranzahl, eine geringe Variabilität innerhalb der Cluster und eine hohe Variabilität zwischen den Clustern ist Voraussetzung für alle weiteren Ansätze, da ohne rein technisch/ mathematisch ausreichend heterogene und in sich homogene Cluster auch keine inhaltlichen Aussagen getroffen werden können.

2.2 Inhaltliche Plausibilität

Die nächste Ebene der Evaluation erfolgt anhand von Expertenmeinungen. Zunächst werden dazu die Clustereigenschaften deskriptiv ausgewertet, um einen ersten Eindruck zu gewinnen,

wie sich die Cluster zusammensetzen. Hierzu können z. B. Mittelwerte, Minima, Maxima, sowie die Quartile numerischer Variablen, oder die empirische Verteilung kategorialer Variablen genutzt werden. Eine weitere Möglichkeit besteht darin, jedes Cluster durch die Beschreibung eines „zentralen Falls“, der sich bezüglich der berücksichtigten Abstandsmaße am wenigsten von den anderen Fällen seines Clusters unterscheidet zu charakterisieren. Auch die Bestimmung der am häufigsten beobachteten Sequenzen von klinischen Ereignissen innerhalb des jeweiligen Clusters kann zur Beschreibung eines Clusters genutzt werden. Des Weiteren wird die Frage gestellt, ob die Merkmale des Clusterings an sich (Anzahl, Größe der Cluster, etc.) aus fachlicher, klinischer Sicht Sinn ergeben. Die Beobachtungen in den gefundenen Clustern werden auf inhaltliche Plausibilität und Ähnlichkeit überprüft. Während die Ähnlichkeit im vorherigen Kapitel rein technisch/ mathematisch beurteilt wurde, soll hier die Beurteilung fachlich/ inhaltlich erfolgen, insbesondere auch, um die Sinnhaftigkeit der für das Clustern benutzten Abstandsmaße, die in den Feedbackrunden zwischen LKR und Offis ausgewählt wurden, anhand des Ergebnisses abschließend zu bewerten. Auch die inhaltliche Ferne der Cluster wird abgeschätzt, um zu beurteilen, ob die Abstände zwischen den Clustern nicht nur technisch, sondern auch inhaltlich Sinn ergeben.

2.3 Methodik Expertenmeinungen

2.3.1 Kontext und Zielsetzung

Zur Evaluation des entwickelten Clustering-Tools wurde eine qualitative Untersuchung durchgeführt. Ziel war es, die Anwendbarkeit, Funktionalität und den Nutzen des Tools in der Praxis zu bewerten. Der Fokus lag darauf, wie gut das Tool die Arbeit mit klinischen Krebsregisterdaten unterstützt, insbesondere bei der Analyse von Patientenclustern. Die Untersuchung wurde mit Expertinnen und Experten aus der klinischen Auswertungsstelle des LKR NRW durchgeführt. Diese Abteilung wurde aufgrund ihrer spezifischen Expertise in der Arbeit mit klinischen Daten und ihrer Zugehörigkeit zur Zielgruppe des Tools ausgewählt.

2.3.2 Auswahl der Teilnehmenden

Für die Untersuchung wurden sechs potenzielle Fachpersonen aus der klinischen Auswertungsstelle kontaktiert. Diese Personen wurden aufgrund ihrer einschlägigen Erfahrung und ihrer Position innerhalb der Zielgruppe des Tools ausgewählt. Nach einer Live-Demonstration erklärten sich drei der Fachpersonen bereit, an der Bewertung teilzunehmen. Die Teilnehmenden wurden bewusst aus dem Team ausgewählt, das die Hauptnutzerinnen und -nutzer des Tools repräsentiert, um praxisrelevante Rückmeldungen zu gewährleisten.

2.3.3 Datenerhebung

Die Datenerhebung erfolgte in Form einer interaktiven Live-Demonstration des Clustering-Tools, gefolgt von einer explorativen, selbstgesteuerten Nutzung des Tools durch die Teilnehmenden. Dabei wurde besonderer Wert auf die Bewertung folgender inhaltlicher Aspekte gelegt:

- Auflösung der gewählten UMAP Projektion
- Die Anzahl und Konfiguration der generierten Patientencluster
- Die medizinische Plausibilität der erstellten Cluster (innerhalb der Cluster und Abgrenzung zwischen Clustern)

- Mögliche Anwendungsfälle
- Feedback bzgl. Anpassungen und Feature-Wünschen

Anstatt alle möglichen Cluster im Detail zu analysieren, hatten die Teilnehmenden die Möglichkeit, sich explorativ auf diejenigen Cluster zu konzentrieren, die sie als besonders relevant erachteten. Dies ermöglichte eine fokussierte und praxisnahe Evaluation der Ergebnisse.

2.3.4 Datenanalyse

Die Rückmeldungen der Teilnehmenden wurden systematisch gesammelt und qualitativ ausgewertet. Dabei wurden zentrale Themen wie Benutzerfreundlichkeit, Funktionalität und die fachliche Validität der Ergebnisse identifiziert. Der explorative Ansatz erlaubte es, spezifische Stärken und Schwächen des Tools hervorzuheben sowie Verbesserungspotenziale zu ermitteln.

2.4 Klinischer und fachlicher Nutzen

Der dritte Aspekt unter dem der Erfolg der Clustering-Modelle betrachtet wurde, ist die Fragestellung, ob sich neben der Bewertung von rein technischen Qualitätskriterien oder einer qualitativen Bewertung der Plausibilität der Cluster, auch in statistischen Auswertungen, die routinemäßig von Krebsregistern durchgeführt werden, relevante Unterschiede zwischen den gefunden Clustern zeigen. Eine häufig auf Krebsregisterdaten angewendete Methode ist die Überlebenszeitanalyse, etwa mit dem Kaplan-Meier Ansatz oder dem relativen Überleben. In unserer Anwendung wollten wir überprüfen, ob durch das Clustern gefundene Behandlungsverläufe einen prognostischen Wert für den Patienten haben, in dem das Überleben nach der Behandlung pro Cluster untersucht wird. Der klinische-fachliche Nutzen wurde hier so definiert, dass sich die gefundenen Cluster auch nach Adjustierung für bekannte Störgrößen wie Diagnosealter, Tumorstadium, Grading oder Tumorhistologie in ihrem 4-Jahres-Überleben unterscheiden, sodass davon ausgegangen werden kann, dass die Behandlungsverläufe, die typisch für das jeweilige Cluster sind, einen eigenen prognostischen Wert haben.

3. Ergebnisse

3.1 Technische Plausibilität

3.1.1 Abstimmbarkeit mit den Originaldaten

Wir haben zunächst überprüft, ob die von OFFIS an das LKR übermittelten Clusterergebnisse mit den Originaldaten abstimmbar sind, ob also die Clusterergebnisse den Originaldaten 1:1 zugeordnet werden können.

Die für die Cluster-Verfahren verwendeten Daten stammen aus einem Datenexport, den das LKR mit verschlüsselter Patienten-ID OFFIS zu Verfügung gestellt hat. Diese Daten wurden mit Zuordnungen zu den jeweiligen Cluster-Verfahren und UMAP-Koordinaten dem LKR zurückgespielt und dort wieder entschlüsselt.

Alle 17.822 Fälle konnten erfolgreich den Originaldaten zugeordnet werden.

3.1.2 Alle Beobachtungen im Ergebnis vorhanden und einem Cluster zuzuordnen

Als nächstes untersuchen wir, ob jede übermittelte Beobachtung in den Clusterverfahren berücksichtigt wurden und einem Cluster zugeordnet wurden.

Aus Performance-Gründen und um eine bessere Übersicht zu gewährleisten wurden nicht alle Daten mit ICD-Code ="C50.XX" für das Clustering berücksichtigt, sondern nur Fälle des Diagnosejahres 2019.

Alle Fälle, die ins Clustering-Modell eingegangen sind, waren auch im Ergebnis vorhanden.

Die Verfahren DBScan und HDBScan lassen eine Nicht-Zuordnung von Verläufen zu einem Cluster technisch zu. Diese Verläufe werden aber als Ausreißer markiert und bilden somit ein weiteres „Ausreißer-Cluster“. Berücksichtigt man dies, konnte jeder Verlauf einem Cluster zugeordnet werden.

Die Ausreißer-Fälle wurden für den DBSCAN beispielhaft untersucht. Es konnte keine besonderen Behandlungsverläufe gefunden werden, die diese Gruppe kennzeichnen, alle Behandlungsarten (OP, Radiotherapie und Systemische Therapie) waren vorhanden.

3.1.3 Parametertuning und Verfahrensauswahl

Cluster-Verfahren haben im Allgemeinen eine Anzahl von Parametern, die man an die Verfahren übergeben muss. Häufig bestimmen diese Parameter über die Größe und Anzahl der gefundenen Cluster. Manche der genutzten Cluster-Verfahren sind sehr Parameter-Tuning intensiv und vor allem auch anfällig für kleine Veränderungen in den Parametern. Darum mussten wir uns für eine Parameterkombination für jedes Clustering-Verfahren entscheiden.

Übergreifend über alle Verfahren muss auch der Parameter n_neighbors des UMAP-Algorithmus bestimmt werden. Kleine Einstellungen dieses Parameters legen höheres Gewicht auf die globale Struktur der Daten. Mit größer werdendem Parameter wird die lokale Struktur um einzelne Punkte herum stärker gewichtet. Höhere Parametereinstellungen bedeuten beim

UMAP-Verfahren also ein stärkeres „Zusammenziehen“ von nahe beieinanderliegenden Punkten. Wir wählen aus vier arbiträr gewählten Parametereinstellungen: 5, 25, 100 und 200. Die folgenden Grafiken zeigen die UMAP-Projektion mit den unterschiedlichen Parametereinstellungen.

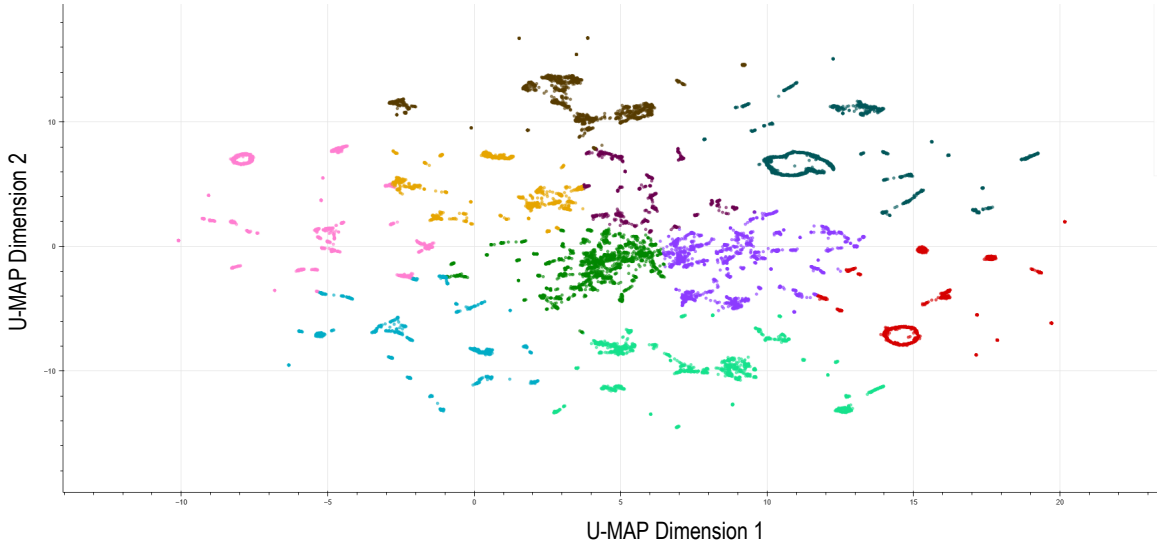


Abbildung 1: Cluster nach UMAP-Projektion mit n_neighbors = 5 und KMedoids (K=10)

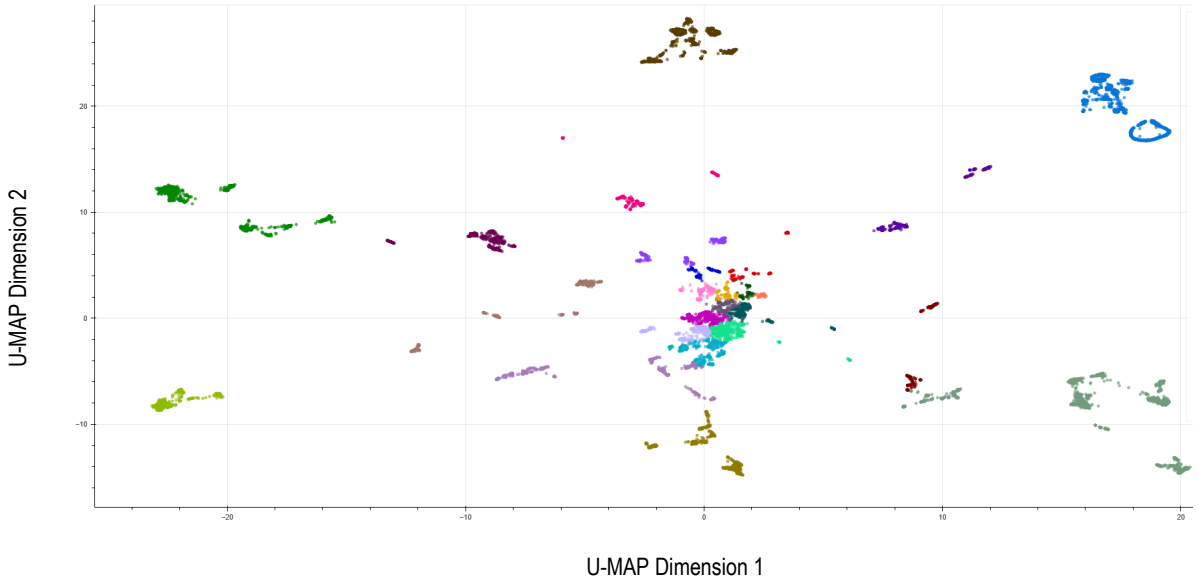


Abbildung 2: Cluster nach UMAP-Projektion mit n_neighbors = 25 und KMedoids (K=25)

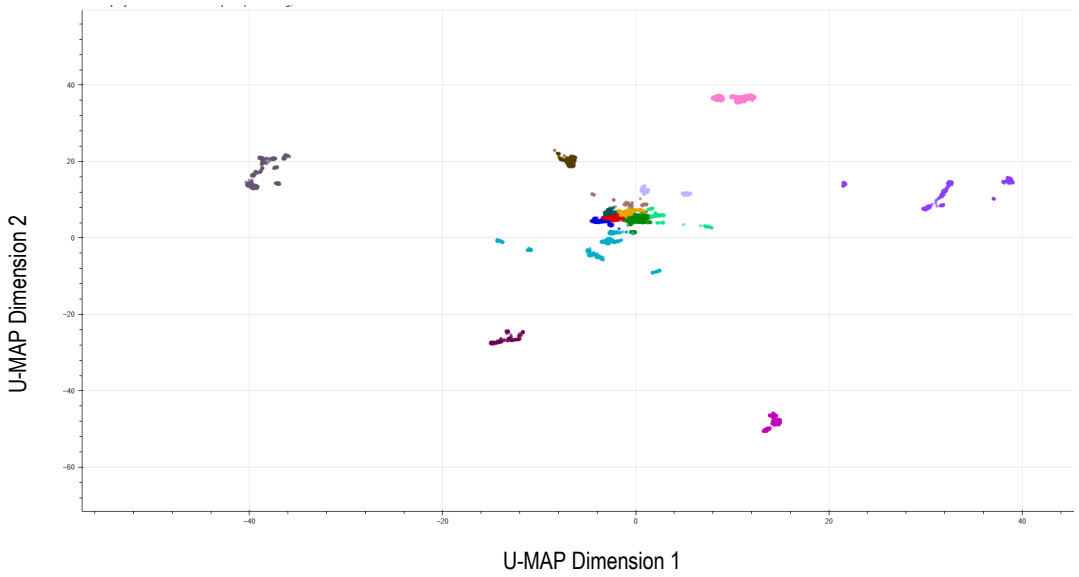


Abbildung 3: Cluster nach UMAP-Projektion mit n_neighbors = 100 und KMedoids (K=15)

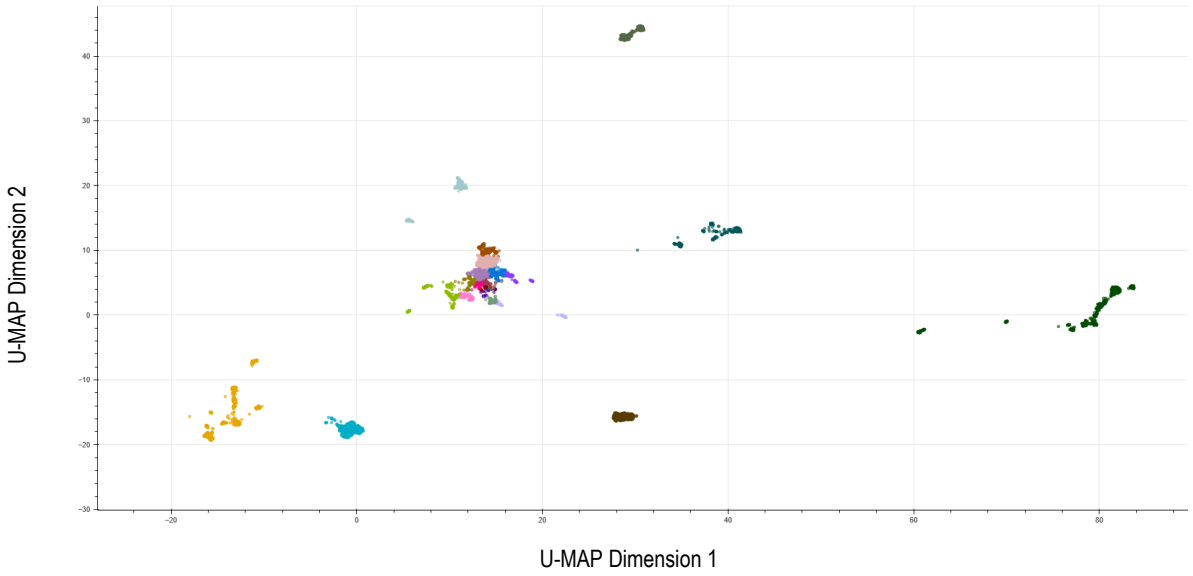


Abbildung 4: Cluster nach UMAP-Projektion mit n_neighbors = 200 und KMedoids (K=30)

Die Farben in den Abbildungen entsprechen den Clustern, die mit dem KMedoids-Algorithmus gefunden wurden, dies ist für die Auswahl des UMAP-Parameters aber nicht relevant.

Man sieht in den Abbildungen, dass mit größer werdendem Parameter mehr Punkte zusammengefasst werden. Für n_neighbors = 5 bilden sich sehr viele kleine und dicht beieinanderliegende Punktwolken. Für n_neighbors = 200 bilden sich nur noch sehr wenige, größere.

Die Auswahl, welche Einstellung wir nehmen sollten, basierte hauptsächlich darauf, in welchem Detailgrad wir in der Lage sein wollen, Unterscheidungen zwischen den Clustern vorzunehmen. Das kann auch von der Anwendung abhängen. Interessieren wir uns für kleine Un-

terschiede zwischen sehr speziellen Behandlungsverläufen, sollte eine kleine Parametereinstellung gewählt werden. Für eine umfassendere, gröbere Unterscheidung ist eine größere Parametereinstellung ausreichend. Da wir keine spezielle Fragestellung untersuchen wollen, sondern eine allgemeine Abschätzung der Nutzbarkeit und Relevanz des Clusterings vornehmen wollen, wurde eine mittlere Parametereinstellung gewählt. Da für Werte zwischen n_neighbors = 100 und n_neighbors = 200 visuell kaum noch Unterschiede erkennbar waren, haben wir 25 als mittlere Einstellung definiert und uns für weitere Auswertungen auf diese Einstellung beschränkt.

Im Folgenden wählen wir Hyperparametereinstellungen für die einzelnen Clusterverfahren.

Zur Beurteilung der Hyperparameter nutzen wir primär den Silhouette-Score. Allgemein ist es bei Cluster-Verfahren erstrebenswert, wenn die gebildeten Cluster möglichst homogen innerhalb der Cluster sind. Gleichzeitig sollten die Cluster untereinander möglichst heterogen sein. Der Silhouette-Score vereint die Homogenität innerhalb der Cluster und die Heterogenität zwischen den Clustern zu einem einzelnen Score und eignet sich also gut für eine Auswahl von Hyperparametereinstellungen.

Theoretisch ist eine automatische Hyperparameteroptimierung möglich, wobei die Einstellung mit dem höchsten Silhouette-Score gewählt wird. Es kommt aber vor, dass stark unterschiedliche Hyperparametereinstellungen (und damit zum Beispiel stark unterschiedliche Clusteranzahlen) zu ähnlichen Silhouette-Scores führen. Für solche Fälle wurde ein manueller Eingriff zugelassen um die Möglichkeit zu haben, auch Parametereinstellungen mit leicht schlechterem Silhouette-Score zu erlauben, wo es inhaltlich sinnvoll erscheint.

Da wir für eine inhaltliche Auswertung keinen Anhaltspunkt zu einer sinnvollen Clusteranzahl hatten und da es für sehr unterschiedliche Clusteranzahlen ähnlich gute Silhouette-Scores gibt, haben wir die Clusteranzahl in drei Bereiche unterteilt, 1-50 Cluster, 51-100 Cluster und 101-200 Cluster. Wir betrachten mehr als 200 Cluster als zu groß, um sinnvolle inhaltliche Auswertungen zu machen. In jedem Bereich wählen wir unter den Verfahren DBScan, HDBScan und Hierarchisches Clustern mit verschiedenen Hyperparametereinstellungen dasjenige mit dem höchsten Silhouette-Score. Außerdem haben wir die Anzahl von als Ausreißer gekennzeichneten Verläufe auf maximal 5% der Gesamtdaten beschränkt. KMedoids wurde als ein mögliches Verfahren in Erwägung gezogen, es hat sich aber herausgestellt, dass das Verfahren insgesamt ungewöhnliche Clusterzuordnungen macht. Die heterogene Datenstruktur scheint es dem Verfahren schwierig zu machen, sinnvolle Cluster zu bilden. Rein optisch bilden sich kleinere und größere Cluster, sowie mehr und weniger gut voneinander abgrenzende. Damit scheint das Verfahren Probleme zu haben. Wir schließen es daher von einer weiteren Betrachtung aus, unabhängig von den Silhouette-Scores.

1-50 Cluster

Tabelle 1: Den besten Silhouette-Score hatte in diesem Fall das hierarchische Clustering mit distance threshold 15.

Verfahren	Parameter 1	Parameter 2	Silhouette-Score	Anzahl Ausreißer	Anteil Ausreißer	Anzahl Cluster
hierarchical	15		0,68	0	0%	12
hierarchical	10		0,67	0	0%	15
hierarchical	7,5		0,66	0	0%	22
hdbscan	46	500	0,65	418	2%	14
hdbscan	56	500	0,65	455	3%	14

51-100 Cluster

Tabelle 2: Für diesen Bereich gab es nur 3 Verfahren und Parameterkombinationen. Den besten Silhouette-Score hat das hierarchische Clustering mit distance threshold 3.

Verfahren	Parameter 1	Parameter 2	Silhouette-Score	Anzahl Ausreißer	Anteil Ausreißer	Anzahl Cluster
hierarchical	3		0,66	0	0%	53
hierarchical	2		0,63	0	0%	80
hdbscan	41	100	0,53	855	5%	51

101-200 Cluster

Tabelle 3: Den besten Silhouette-Score hatte in diesem Fall das hierarchische Clustering mit distance threshold 1.

Verfahren	Parameter 1	Parameter 2	Silhouette-Score	Anzahl Ausreißer	Anteil Ausreißer	Anzahl Cluster
hierarchical	1		0,61	0	0%	174
dbscan	0,2	17	0,57	530	3%	125
dbscan	0,2	18	0,56	596	3%	123
dbscan	0,2	19	0,56	642	4%	123
dbscan	0,2	20	0,56	722	4%	121

Zusammenfassung der Verfahrensauswahl

Für die drei Bereiche haben sich die folgenden Verfahren und Parametereinstellungen als die besten im Sinne des Silhouette-Scores erwiesen:

- 1-50 Cluster: Hierarchisch, Distance Threshold 15, 12 Cluster
- 51-100 Cluster: Hierarchisch, Distance Threshold 3, 53 Cluster
- 101-200 Cluster: Hierarchisch, Distance Threshold 1, 174 Cluster

Grundsätzlich hatten kleinere Clusteranzahlen grundsätzlich höhere Silhouette-Scores.

In allen drei Fällen hatte das hierarchische Clusterverfahren die besten Silhouette-Scores, also beschränken wir uns in der weiteren, inhaltlichen Betrachtung auf diese.

3.1.4 Gesamtbeurteilung der technischen Plausibilität

Abhängig vom Verfahren und den Parametereinstellungen scheint es sowohl für kleinere Clusteranzahlen von 1-50 als auch für größere Anzahlen über 100 Ergebnisse mit guten Silhouette-Scores zu geben. Eine rein technische Beurteilung einer optimalen Anzahl konnten wir hier nicht treffen. Wir haben deshalb die drei gefundenen Verfahren im Weiteren inhaltlich betrachtet

3.2 Deskriptive Auswertung und Inhaltliche Beurteilung der Clustereigenschaften

Abbildung 5 zeigt die UMAP-Projektion eingefärbt in den Clustern der Methode mit dem höchsten Distance Threshold (TSH) = 15 und der geringsten Clusterzahl. Es gibt zwölf von der Größe her sehr unterschiedliche Cluster (c 0 - c 11), die größtenteils auch optisch in der UMAP-Projektion den dort gebildeten Punktewolken entsprechen, insbesondere am Rand.

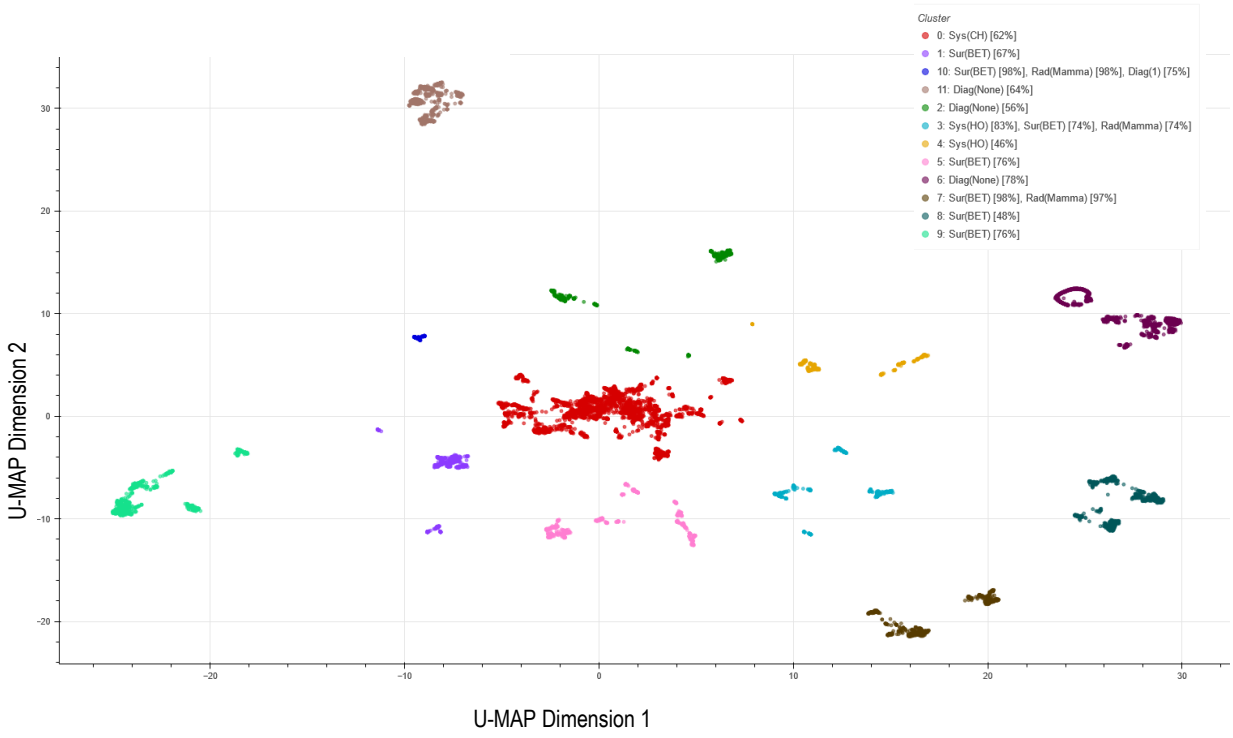


Abbildung 5: UMAP-Projektion mit TSH=15

Ebenfalls in Abbildung 5 (links oben) zu sehen ist das häufigste Ereignis, das diese Gruppe kennzeichnet, sowie die Angabe, wie viele der Fälle im Cluster dieses Ereignis mindestens einmal enthalten. Bei Cluster 0 zeigt sich, dass nur bei etwas mehr als der Hälfte der fast 5000 Fälle das häufigste Ereignis - Chemotherapie – vertreten ist. Es gibt weitere Cluster mit ähnlich niedrigen Werten, die auf eine hohe Heterogenität innerhalb der Cluster hinweisen (Cluster 2, 4, 8).

Generell ist anhand der Verteilung der einzelnen Fall-Punkte zu erkennen, dass Verläufe mit vielen Ereignissen, insbesondere auch ein Großteil der mit systemischer Therapie (Sys) sich in der Mitte der Projektion befinden (TSH=15, C 0), während sich an den äußersten Rändern solche finden, die nur das Diagnose-Ereignis haben, unterteilt nach Altersgruppen 50-69 (TSH=15, c 11) und >69 (TSH=15, c 6).

Das Cluster 11 (TSH=15) enthält Fälle mit Diagnosealter 50-69 ohne weitere Behandlung, die in der nächst höheren Auflösung in zwei Cluster (TSH=3, c 0 und c 39) aufgeteilt werden (siehe Abbildung 6). Auf den ersten Blick scheint der Grund dafür das Grading von 2 bzw. 3 zu sein (Medoid und Tooltip).

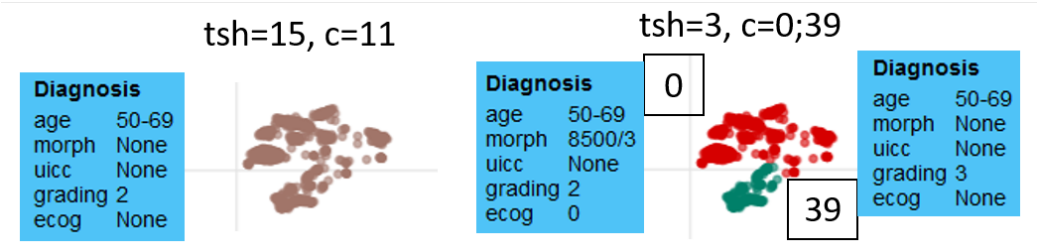


Abbildung 6: Cluster 11 (TSH=15, links) und Cluster 0 und 39 (TSH=3) zusammen mit dem jeweiligen Medoid in der UMAP-Projektion

Auch das Cluster 6 (TSH=15), das die Fälle mit Diagnosealter >69 ohne weitere Behandlung enthält (siehe Abbildung 7), wird in der nächst höheren Auflösung in mehrere Cluster (TSH=3, c 20, c 31, c 49 und c 50) aufgeteilt. Auch in dieser Altersgruppe scheint das Grading dafür ausschlaggebend zu sein, was Sinn ergibt, aber auch eine Folge davon sein könnte, dass bei Fällen mit gleichem Grading auch einige andere Merkmale ähnlich sein könnten.

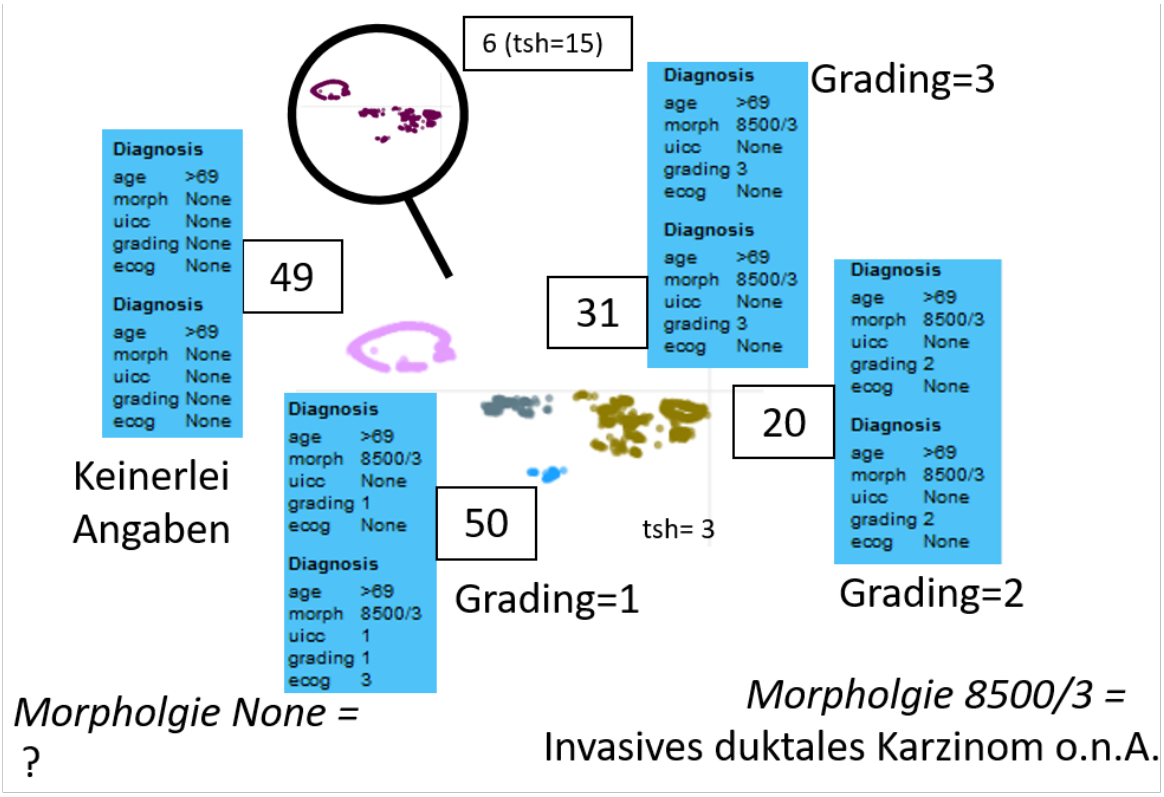


Abbildung 7: Cluster 20, 31 49, 50 (TSH=3) mit dem jeweiligen Medoid in der UMAP-Projektion

Interessant ist hier, dass die Punktewolke, die die dritte Altersgruppe der Fälle ohne weitere Behandlung enthält (<50), bei TSH=15 mit einigen anderen Wolken zu Cluster 2 zusammengefasst wird. In Abbildung 8 sieht man den Medoid, sowie die jeweils drei Fälle, welche diesem am ähnlichsten bzw. am unähnlichsten sind. Dabei sieht man einerseits, dass der Medoid und die ihm am ähnlichsten Fälle Diagnosealter < 50 und nur das Diagnoseereignis haben, andererseits aber auch, dass es Fälle gibt, die stark davon abweichen und ein oder mehrere Therapien hatten. Anders als bei den beiden anderen Nur-Diagnose-Wolken erfolgt hier die Einteilung in Einzelcluster erst in der nächst höheren Auflösung (TSH=3, Abbildung 9). Hier erkennt

man das Cluster Fälle ohne weitere Behandlung <50 (c 28), ein Cluster mit Fällen, die nur Sys erhalten haben (c3), ein Cluster mit OP vor der Diagnose (c 48) und ein typisches den Leitlinien entsprechendes BET + Radiatio (Rad) Cluster (c 29).

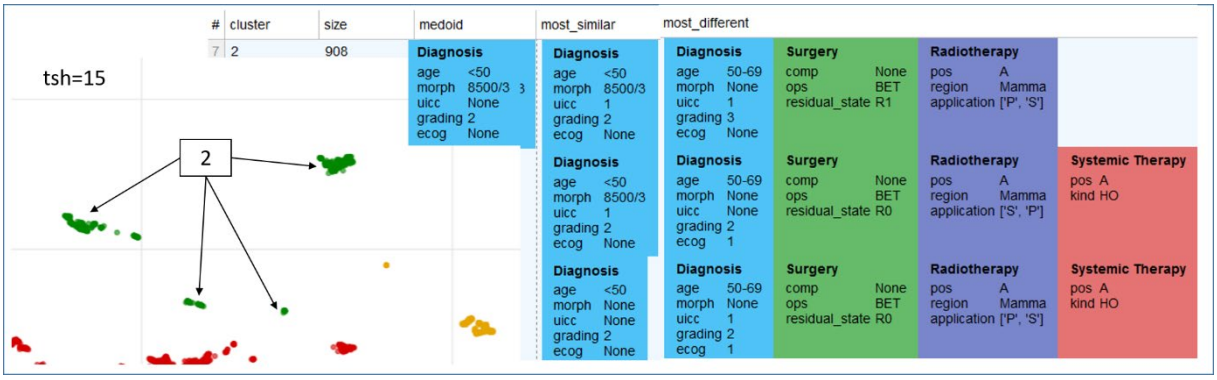


Abbildung 8: Cluster 2 (TSH=15, links) in der UMAP-Projektion und Cluster Informationen: Medoid, ähnlichster und unähnlichster Fall

Die Abbildung 9 zeigt jeweils den Medoid des jeweiligen Clusters. Ein Medoid ist ein repräsentativer Fall eines Clusters, dessen Abstand zu allen anderen Fällen desselben Clusters in der Summe minimal ist. Der Medoid ist jedoch nicht unbedingt intuitiv für die menschliche Logik begreifbar, da z.B. bei Cluster 28 der Medoid eine OP vor Diagnose hat. Beim Tooltip aber erkennt man, dass nur wenige Ausnahmen in diesem Cluster ein OP-Ereignis haben und diese ebenso wie der Medoid keine Angaben zur OP enthalten, während die große Mehrheit nur ein Diagnoseereignis hat. Bei Cluster 3 z.B. wird die Therapieart der einen Sys des Medoid mit CI (=Chemo- + Immun-/Antikörpertherapie) angegeben, jedoch macht haben nur 8,5% aller Fälle eine Sys dieser Art, während 88,7% eine Chemotherapie erhielten (=CH). Der Medoid ist daher nicht so zu verstehen, dass er in allen seinen Merkmalen der Mehrheit aller anderen Fälle entspricht.

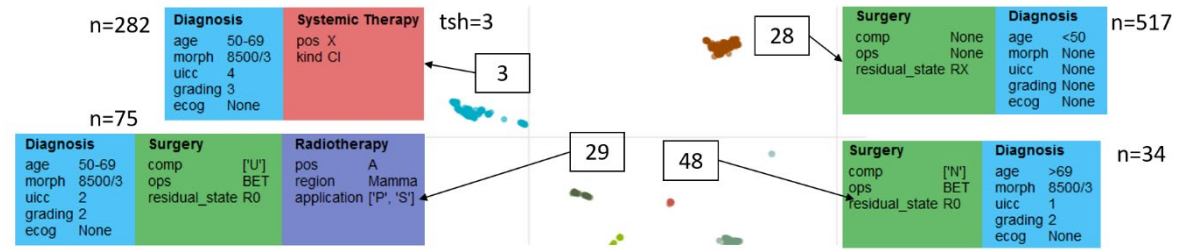


Abbildung 9: Cluster 3, 29, 28, 48 (TSH=3) mit dem jeweiligen Medoid in der UMAP-Projektion

Hier ergibt die Aufteilung des TSH=15 Clusters 2 auf mehrere Cluster inklusive des Randclusters 28 mit nur-Diagnose-Fällen in der nächst feineren Auflösung Sinn.

Schaut man sich nun die höchste Auflösung der Cluster (TSH=1) an (siehe Abbildung 6), so erkennt man auch hier eine weitere Aufteilung des Clusters 28 (TSH=3) mit Fällen ohne weitere Behandlung (<50) in drei Cluster mit unterschiedlichen Grading-Angaben (TSH=1, c 9, c 121, c 147). Die Cluster 89 und 48 (TSH=3) bleiben in der höheren Auflösung (TSH=1) erhalten (c 59, c 97). Das zweitgrößte Untercluster des Cluster 2 (TSH=15), Cluster 3 (TSH=3), wird bei TSH=1 weiter aufgeteilt in drei Cluster. In Abbildung 10 sieht man auch den entsprechenden Medoid zu den Clustern, der jedoch auch hier nicht bei allen Clustern eine gute Übersicht über

den Inhalt gibt. Bei Cluster 13 (TSH=1) zeigt das Tooltip, dass die Mehrheit der Fälle der Altersgruppe 50-69 angehören und Chemotherapie als einzige Therapie erhalten haben, während bei Cluster 27 alle Altersgruppen und bei Cluster 56 Fälle mit nur Diagnosealter <50 vertreten sind.

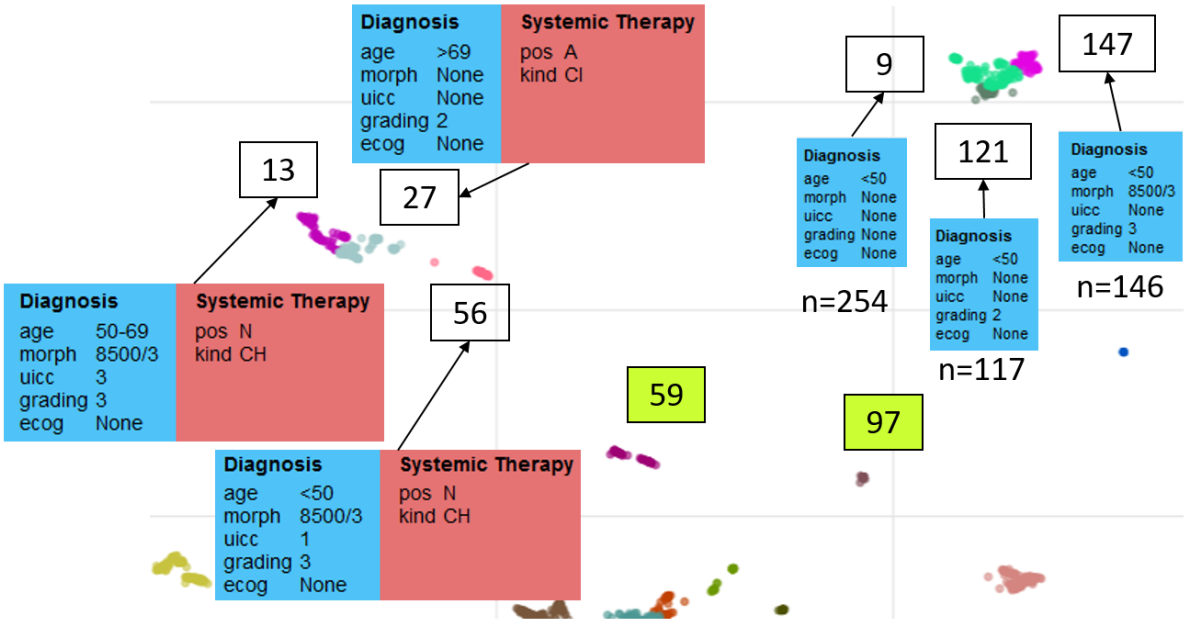
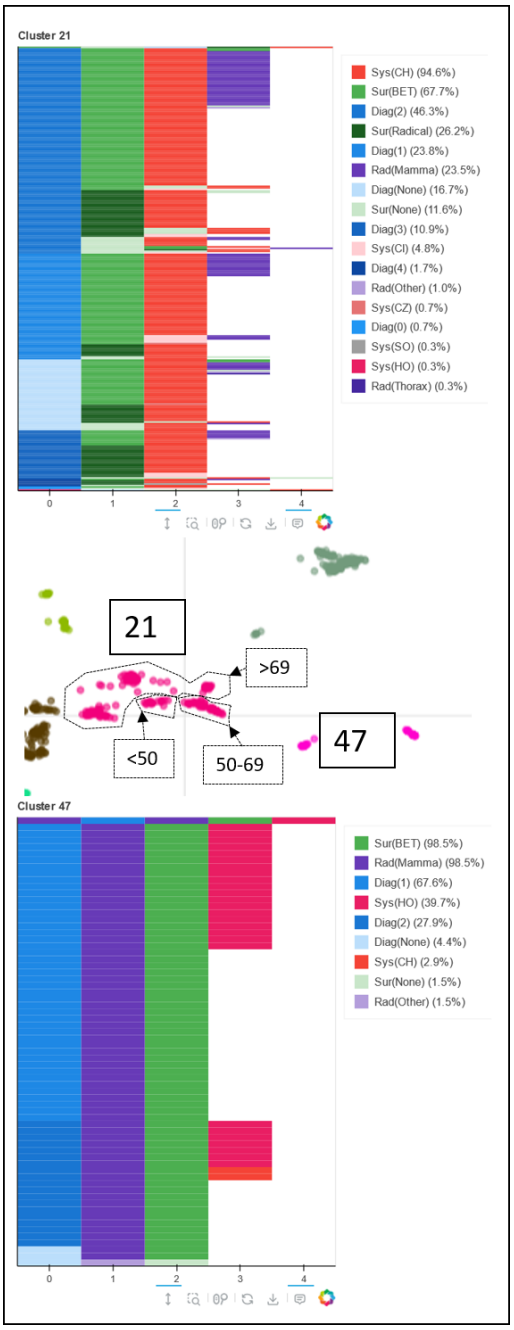


Abbildung 10: Cluster 13, 27, und 56 zusammen mit dem jeweiligen Medoid, Cluster 9, 121 und Cluster 147 zusammen mit dem jeweiligen Medoid und der Clustergröße n in der UMAP-Projektion (TSH=1)

Das größte Cluster 0 (TSH=15) enthält mehr als ein Viertel aller Fälle (28%) und fällt auch durch seine Lage im Zentrum der Projektion auf. Beim Tooltip erkennt man, dass sich alle möglichen Ereignis-Kombinationen (OP, Sys, Rad) in diesem Cluster befinden und sich hier auch die Fälle mit langem Therapieverläufen (4 und mehr Ereignisse) befinden. Um einzelne Behandlungsmuster zu erkennen, muss man in der Auflösung einen oder zwei Schritte nach oben gehen und sich die Cluster mit TSH=3 und TSH=1 anschauen.

In Abbildung 11 sind Behandlungssequenzansichten zu sehen. Von links nach rechts sieht man die Art des Ereignisses in chronologischer Reihenfolge beginnend mit dem ersten Ereignis (hier: Diagnose). Jeder Balken entspricht einem Brustkrebs-Fall, jedes Feld einem Ereignis. Die Zahl ganz unten bezeichnet die Position des Ereignisses. In der Legende (rechts neben der Behandlungssequenzansicht) sieht man die Eigenschaften, die ein solches Ereignis haben kann. In der Behandlungssequenzansicht des Clusters 21 (TSH=3, siehe Abbildung 11) sieht man sehr schön, dass hier hauptsächlich Fälle mit der Behandlung OP, gefolgt von Chemotherapie und in einigen Fällen auch noch Rad zusammengefasst sind. Beim Tooltip kann man erkennen, dass die Punktewolken unterteilt nach Alter in der UMAP-Projektion dargestellt werden.



In der noch feineren Auflösung wird Cluster 21 (TSH=3) auf drei Cluster aufgeteilt (TSH=1 c 11, c 50, c 144, siehe Abbildung 8) – allerdings nicht nach Alter, sondern nach Behandlungsmuster. Das ist an dieser Stelle klinisch logisch und daher wünschenswert; es ist auch ein Indiz dafür, dass wir Ähnlichkeiten und deren Gewichtung richtig gewählt haben. Das Modell mit TSH=1 sammelt in Cluster 50 und 144 nur Fälle mit BET und Chemotherapie, in 144 ist zusätzlich noch eine Rad dabei. Cluster 11 hingegen enthält mehrheitlich eine radikale OP und Chemotherapie, jedoch keine Bestrahlung. In allen drei Clustern, aber vermehrt in Cluster 11 finden sich interessante "Ausreißer", die mehr Ereignisse als die anderen Fälle haben oder bei denen die Reihenfolge vertauscht ist. So ist in Cluster 11 z.B. ein Fall mit Sys vor Diagnose zu finden, was implausibel ist.

In der Nähe von Cluster 21 (TSH=3) findet sich noch ein weiteres, kleineres Cluster, c 47, das bei TSH=15 ebenfalls noch zu Cluster 0 gehört und in der nächst höheren Auflösung auf zwei Cluster (TSH=1, c 136 und c 137) aufgeteilt wird. In der Behandlungssequenzansicht sieht man auch hier die Aufteilung von einem Mix aus Behandlungsmustern (TSH=3, c 47) in zwei klar getrennte Behandlungsstränge, nämlich Rad gefolgt von BET (c 137) und Rad gefolgt von BET und Hormonbehandlung (c 136). Beide entsprechen nicht den Leitlinien, welche BET gefolgt von Rad vorsehen und regen zu weiteren Nachforschungen an.

Abbildung 11: Darstellung von Cluster 21 und 47 in der U-MAP-Projektion (TSH = 3) und deren Behandlungssequenzansicht

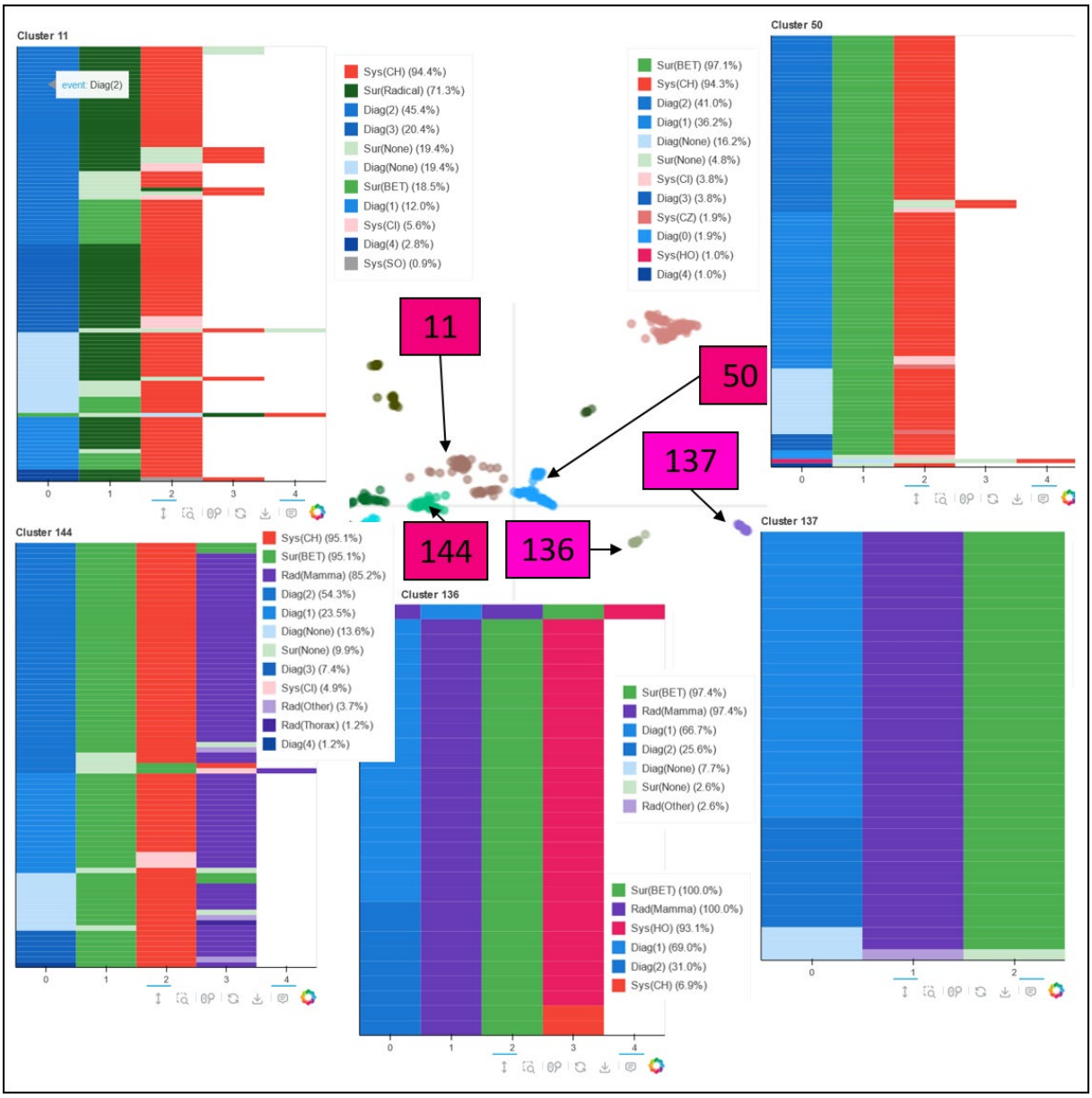


Abbildung 12: Cluster 11, 50, 144 und Cluster 136, 137 in der UMAP-Projektion (TSH=1), sowie deren Behandlungssequenzansichten

Cluster 120 (TSH=1, Teil von TSH=15, c 0) ist ein typisches Cluster aus der Mitte der UMAP-Projektion. Die Behandlungssequenzansicht zeigt eine Fülle von unterschiedlichen Behandlungsmethoden und deren Ausprägungen und auch die Lage (siehe Abbildung 13, gezeigt sind zwei verschiedene Zoomansichten) im Zentrum zeigt, dass dieses Cluster räumlich ebenso schwierig abzugrenzen ist von der Umgebung, wie inhaltlich. Die einzige Gemeinsamkeit scheinen hier multiple verschiedene Sys zu sein, vermischt mit einzelnen Ops und Rads. Hier wäre es sicherlich interessant, genauer zu analysieren, ob der Algorithmus hier tatsächlich einen inhaltlich relevanten neuen Behandlungsverlauf entdeckt hat, oder in dieses Cluster einfach nur "undefinierbare Reste" gepackt wurden, oder ob dieser Wust von verschiedenen Sys vielleicht eher auf implausible Daten hinweist.

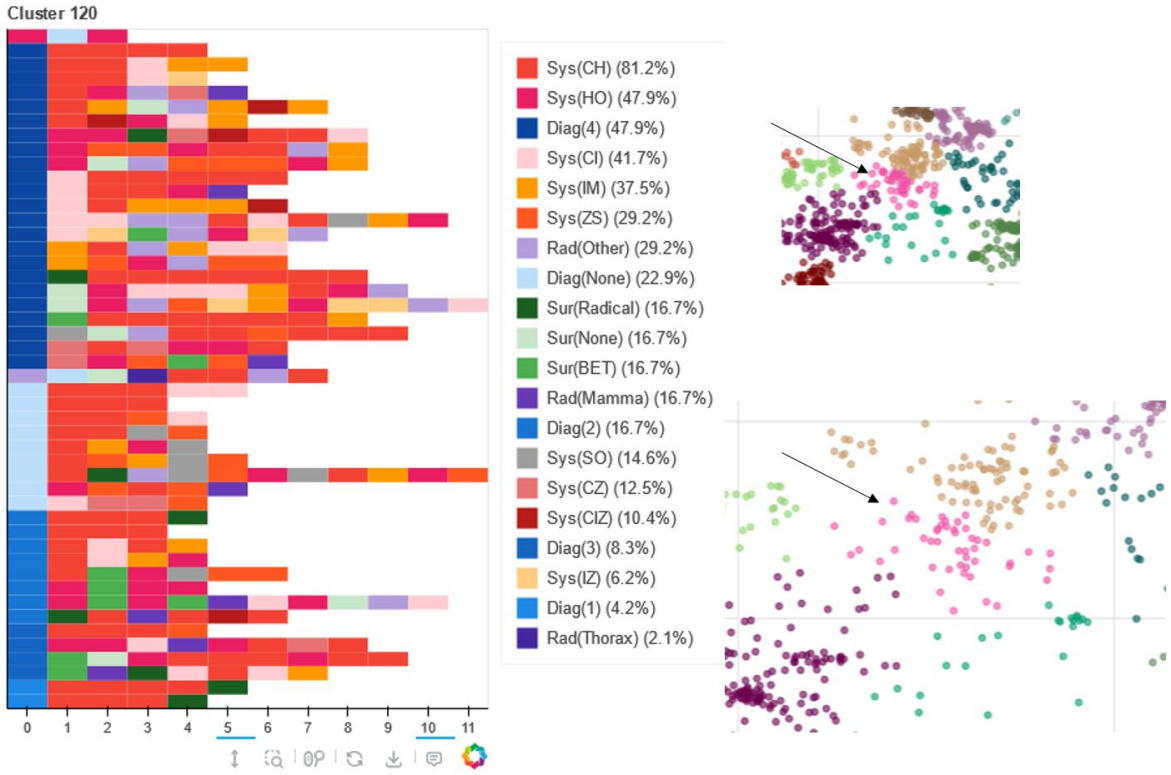


Abbildung 13: Cluster 120 in der UMAP-Projektion (TSH=1, rechts) in zwei verschiedenen Zoom-Ansichten, sowie dessen Behandlungssequenzansichten (links)

Betrachtet man Cluster 17 (TSH=3, Abbildung 11, olivgrün, rechts oben), welches ebenfalls noch zu Cluster 0 (TSH=15) gehört, genauer, so stellt man den gleichen Effekt fest, wie bereits beschrieben (Abbildung 14): Die Untercluster der zentralen großen Clusters 0 (TSH=15) zeigen erst in der höchsten Auflösung (TSH=1) logisch nachvollziehbare Behandlungsgruppen an, hier erfolgt eine Trennung in OP (95% Radical), gefolgt von Bestrahlung (c 72) und nur Bestrahlung (c 155), aber mit Angabe von Position zur OP, was darauf schließen lässt, dass hier OP-Meldungen fehlen.

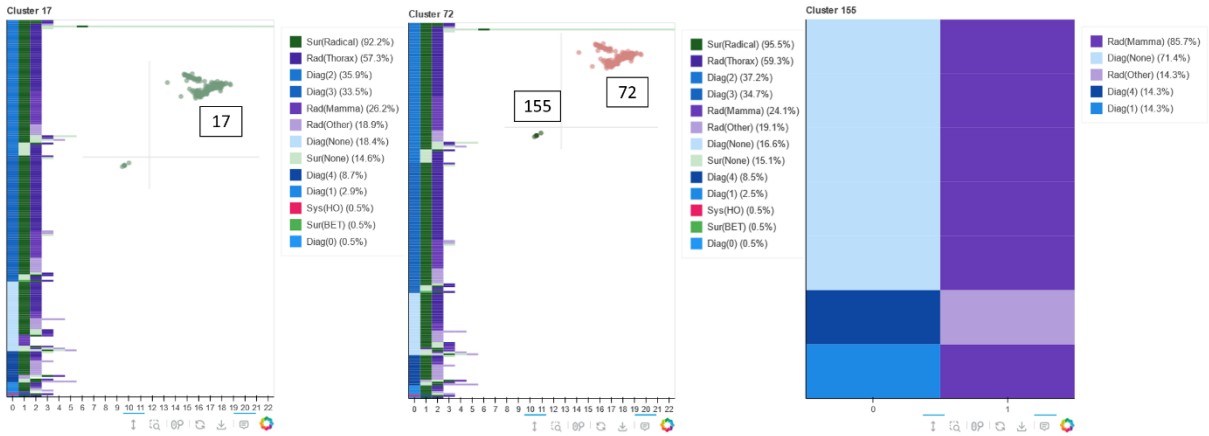


Abbildung 14: Cluster 17 (TSH= 3) und Cluster 72, 155 (TSH=1) in der UMAP-Projektion, sowie deren Behandlungssequenzansichten

Die den Leitlinien entsprechende Behandlung mit BET, gefolgt von Rad (siehe auch Mamma QI 8) findet sich in mehreren Clustern wieder, teilweise kombiniert mit anderen Behandlungsarten, wie CH oder HO Behandlung oder weiteren OPs (z.B. TSH=3, c 10, c 11, c19).

Besonders hervor tritt Cluster 27 (TSH=3 ~ TSH=1, c 111, siehe Abbildung 15) sowohl durch seine Lage weit außen am Rand und geringe Größe (n=6), als auch durch seine Homogenität. Alle in der Summary dargestellten Eigenschaften der Behandlungsverläufe sind identisch, der Medoid gleicht den 5 ähnlichsten Fällen zu 100% und diese sind wiederum gleichzeitig auch die fünf unähnlichsten – es scheint als hätte der Algorithmus hier 6 nahezu identische Behandlungsverläufe in ein Minicluster gepackt – und das bereits in der ersten höheren Auflösung (TSH=3). Ein Grund dafür könnte deren Altersgruppe >69 sein.

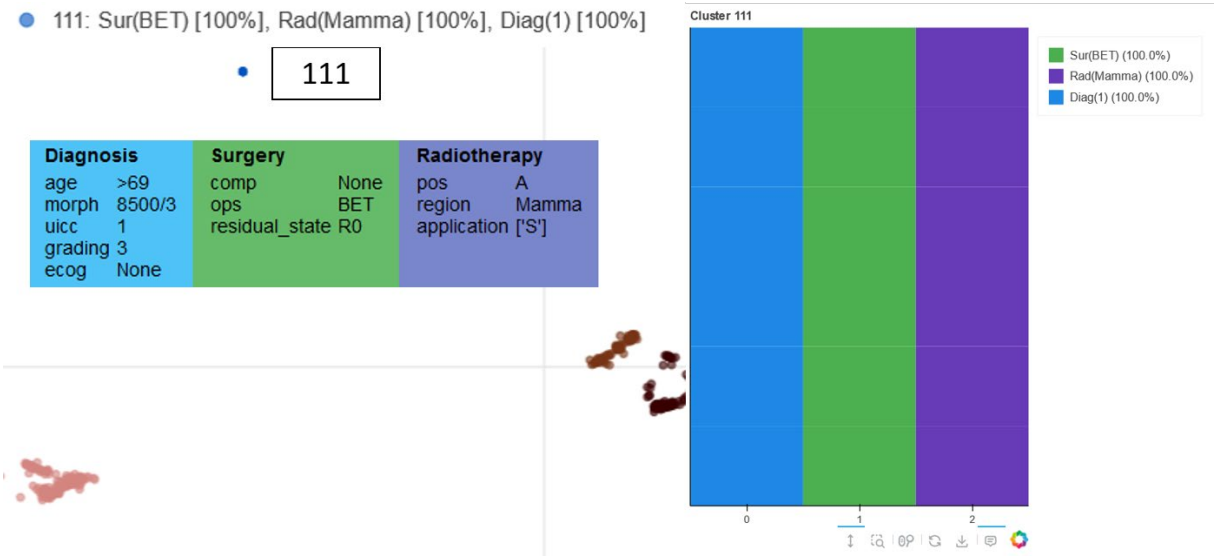


Abbildung 15: Cluster 111 in der UMAP-Projektion (TSH=1), sowie dessen Medoid und Behandlungssequenzansicht

Auch in Cluster 5 findet sich dieses Behandlungsmuster. In der geringsten Auflösung (siehe Abbildung 16) zeigt der Medoid auch exakt dieses. Doch erkennt man bereits rein optisch (von links nach rechts) folgende fünf Untergruppen: Nur OP, OP-Radiatio-(weitere Therapien), OP-Systemische Therapie, OP-Sys-Rad, OP-Sys-Rad-weitere Therapien, respektive Cluster 25, 33, 42, 38 und 5 mit TSH= 3. Abbildung 18 zeigt diese fünf Cluster mit ihrem jeweiligen Medoid (Teil A) und

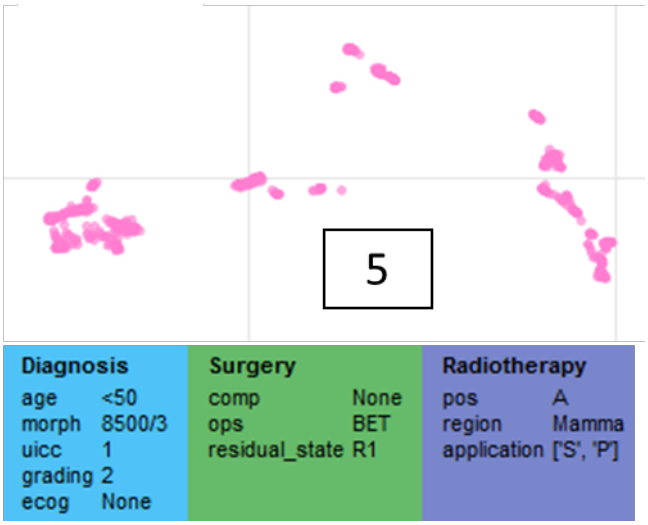


Abbildung 16: Cluster 5 in der UMAP-Projektion (TSH=15), sowie dessen Medoid

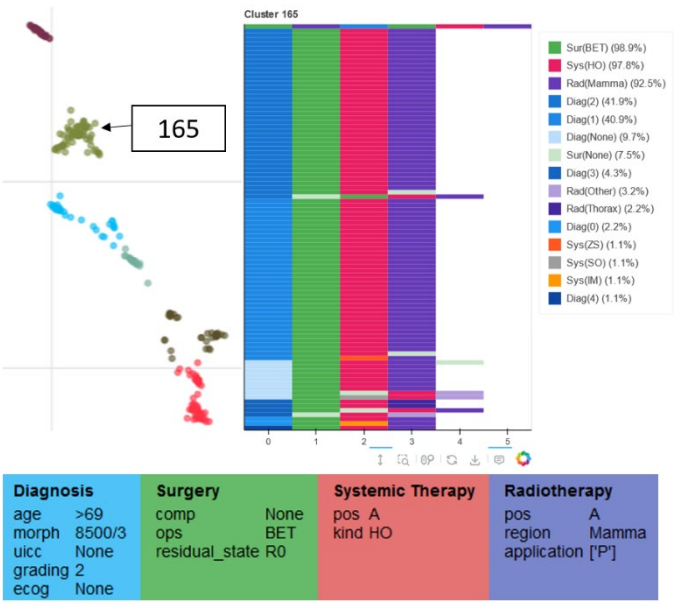


Abbildung 17: Cluster 165 in der UMAP-Projektion (TSH=1), sowie dessen Medoid

Cluster 25 in der höchsten Auflösung (TSH=1). Das Cluster 25 (TSH=3) wird auf 3 Cluster verteilt, in Ermangelung eines Medoiden für einzelne Clusteruntergruppen werden jeweils drei Fall-Summaries aus einer Punktewolke angezeigt. So sieht man z.B., dass Cluster 10 (OP=Radical) auch einen optisch abgegrenzten Clusteranteil mit HO enthält. Erwähnenswert ist hier auch, dass die OP-Methoden BET/ Radical oder keine der beiden (None) sich wie zwei Linien durch die UMAP-Projektion (angedeutet durch die gestrichelte rote Linie) und auch durch Cluster 16 ziehen. Offenbar waren für Cluster 16 andere Eigenschaften wichtiger als die OP-Methoden.

Auch Cluster 38 (TSH=3, siehe oben und Abbildungen 13 und 14) enthüllt in der nächst höheren Auflösung ein sehr interessantes Untercluster (TSH=1, c 165, siehe Abbildung 17). Hier finden wir ähnlich wie im Minicluster 111 (bzw. TSH=3, 27) eine eher kleine Anzahl (n=93) Fälle mit Diagnosealter >69 und einem typischen LL-Verlauf: BET, gefolgt von Rad. Allerdings haben fast 100% dieser Fälle auch noch eine HO im Verlauf. Räumlich in der UMAP-Projektion sind sie weit von Cluster 111 entfernt.

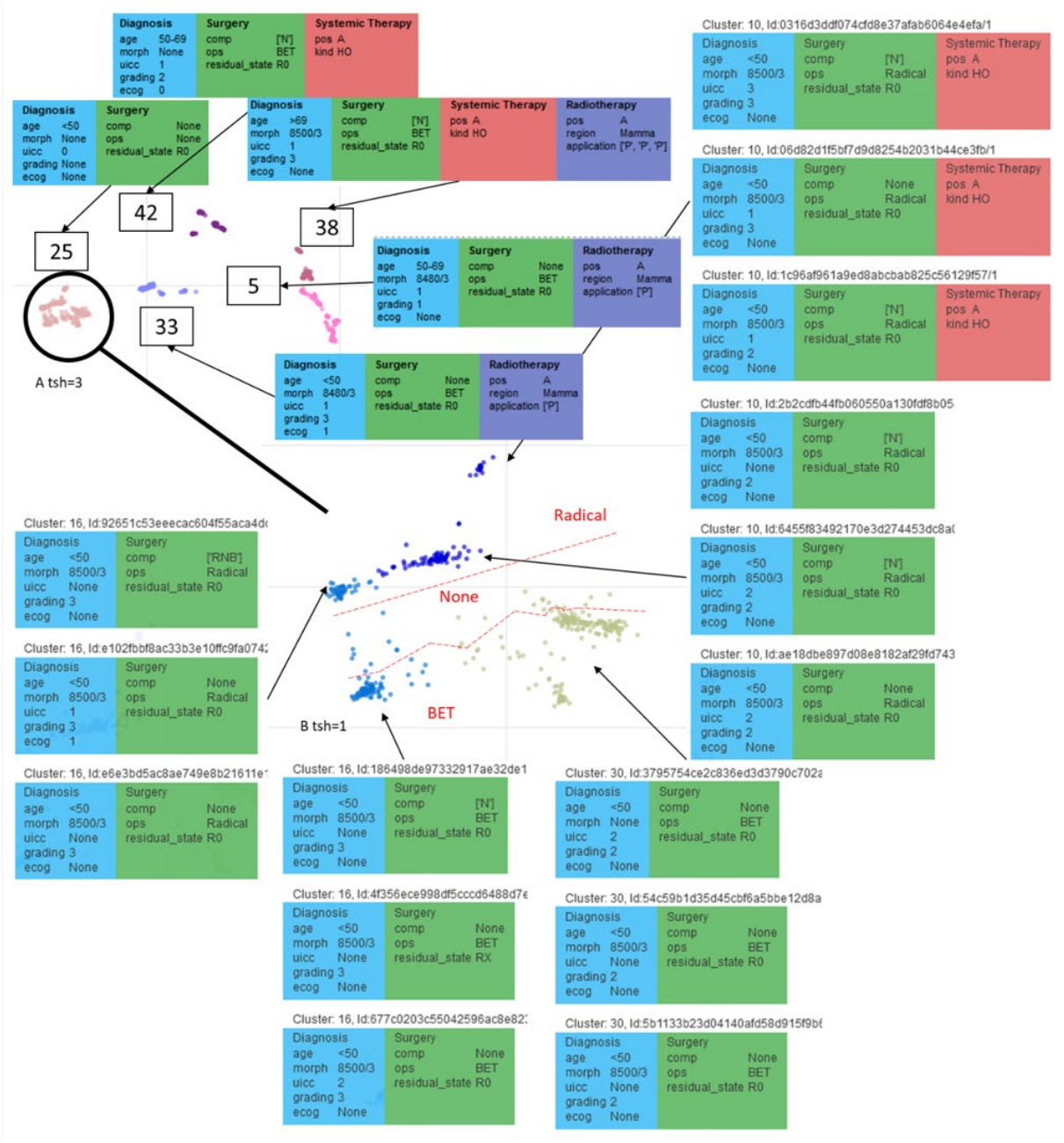


Abbildung 18: Teil A: Cluster 5, 25, 33, 38, 42 in der UMAP-Projektion (TSH=3), sowie deren Medoid, Teil B: Cluster 25 in der feineren Granulierung (TSH=1) und höheren Auflösung in der UMAP-Projektion, manuell unterteilt in Untergruppen sowie deren typische Behandlungsverläufe

Das größte BET + Rad Cluster ist Cluster 7 (TSH=15, n=1176), ein sehr großes homogen scheinendes Cluster. Selbst die ähnlichsten und unähnlichsten Fälle unterscheiden sich kaum, offensichtlich nur in Bezug auf das Alter. In der nächst höheren Auflösung (TSH=3) kann eine Aufteilung nach Alter und Grading beobachtet werden: >69 (c 45) und 50-69, Grading=1 (c 2) und Grading=2 (c 46).

Dieses Clustering-Verfahren konnte uns außerdem mehrere Cluster mit offenbar implausiblen Daten zeigen.
Als Beispiel sei das Cluster 28 (TSH=3, siehe auch oben) genannt. Es ist äußerst unwahrscheinlich, dass Patienten unter 50 Jahren nicht in irgendeiner Form behandelt werden. Es ist zu

vermuten, dass bei diesen Fällen die entsprechenden Behandlungsdaten nicht ans Krebsregister übermittelt wurden. Ein weiteres Indiz dafür ist, dass der Medoid des größten Clusters in der nächst feineren Auflösung (TSH=1, c 9, 49% von TSH=3, c 28) gar keine der angezeigten weiteren Informationen zur Diagnose enthält, was auf eine schlechte Meldungsqualität zu diesen Fällen hindeutet.

3.3 Expertenmeinung

In einem 90-minütigen Wechsel aus Gespräch und Interview wurden drei Mitarbeiterinnen aus der klinischen Auswertungsstelle des LKR zu den Clusterverläufen befragt. Ihnen wurden vorab die zum Verständnis und zur Interpretation der Ergebnisse nötigen Grundlagen erläutert und dann die Analyse-Dashboards und die Behandlungssequenzansicht vorgestellt.

Der Schwerpunkt dieses Gesprächs richtete sich auf die inhaltlichen Aspekte der Clusterbeurteilung.

3.3.1 UMAP-Projektion

Die UMAP-Projektion wurde als sehr gut geeignet für die visuelle Darstellung des Clustering-Ergebnisses empfunden. Sie veranschaulicht deutlich Clustergröße und –zahl, auch die Abstände der Cluster untereinander erscheinen im Großen und Ganzen sinnvoll. Lediglich die Abstände der Gruppen ohne Therapien wurden als zu groß zueinander empfunden (siehe Abbildung 19).

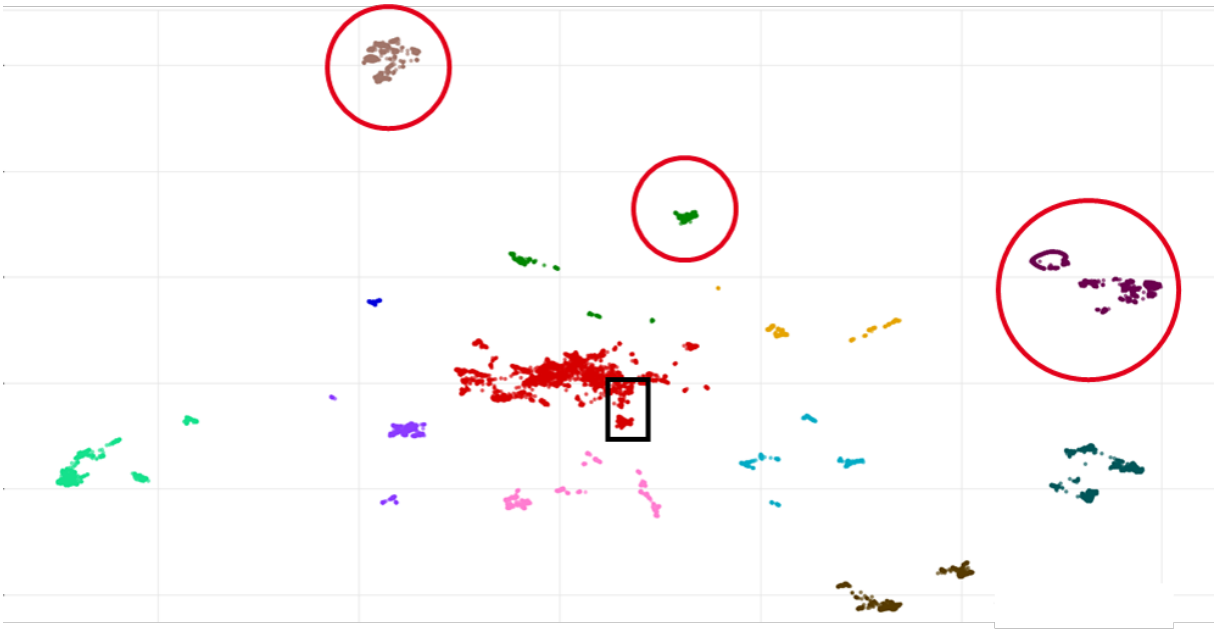


Abbildung 19: UMAP-Projektion mitTSH=15

Die roten Kreise zeigen die Cluster mit *Nur-Diagnose ohne weitere Therapie*.

3.3.2 Clusteranzahl

Als geeignet für weitere inhaltliche Analysen wurde das Model mit 53 Clustern (TSH=3) befunden. Außerdem kam der Vorschlag alle drei hierarchischen Clusterverfahren zu nutzen: Im Verfahren mit wenigen, gröberen Clustern könnte man sich ein interessantes Cluster heraussuchen und dieses in den Verfahren mit der höheren Clusteranzahl suchen, um so eine feinere Aufteilung der Behandlungsverläufe zu analysieren.

3.3.3 Plausibilität zwischen und innerhalb der Cluster

Zur Beurteilung der Plausibilität der Cluster wurde bei der Sichtung der verschiedenen Hierarchischen Cluster Dashboards in viele Cluster genauer geschaut (per Zoomen und Tooltip) und bei den meisten sowohl die Abgrenzung als Cluster als auch die Homogenität innerhalb der Cluster als plausibel befunden. Ein Beispiel dafür sind die Cluster 1 und 32 (TSH 3). Beide enthalten nur Fälle mit einer OP zusätzlich zur Diagnose, Cluster 1 hauptsächlich radikale OPs (87%), Cluster 32 dagegen brusterhaltende OPS (94%).

Es gab jedoch auch Beispiele, wo der Sinn der Aufteilung in die Cluster anhand einfacher Merkmale nicht direkt erfassbar zu sein schien.

Ein exemplarisches Beispiel von benachbarten Clustern wurde daher noch intensiver betrachtet. Es ist im Clusterverfahren mit Threshold=15 ein Teil des Riesenclusters 0 (siehe Abbildung 19, schwarzer Kasten und Abbildung 20), bei den beiden anderen Clusterverfahren wurden dieses auf mehrere Cluster aufgeteilt (siehe Abbildung 20).

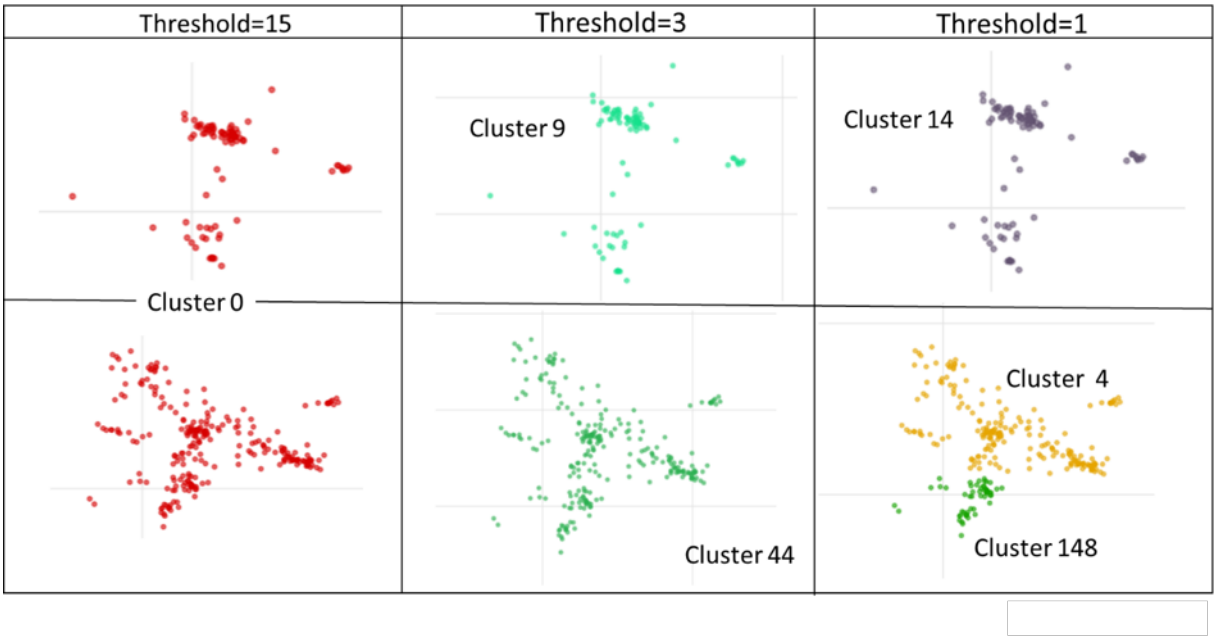


Abbildung 20: ein Teil des Clusters 0 (TSH=15) in der gröbsten (TSH=15) Ansicht und den beiden feineren (TSH=3 und TSH=1)

Cluster 9 (TSH 3), bzw. 14 (TSH 1) wird in den beiden Modellen als ein Cluster dargestellt. Rein visuell wurde der räumliche Abstand zwischen der oberen und der unteren Punktewolke innerhalb des Clusters als sehr groß beurteilt, hier scheint eine Aufteilung auf zwei Cluster besser. Unterstützt wurde dieses Urteil durch die Analyse der Behandlungssequenzansicht (siehe Abbildung 21): Während im oberen Bereich des Clusters 9 bzw. 14 klassische Behandlungsverläufe (mehrere OPs, gefolgt von Rad und Sys) vorlagen, waren die Behandlungsverläufe im unteren Bereich alle eher atypisch bis unplausibel (mind. eine Behandlung vor der Diagnose).

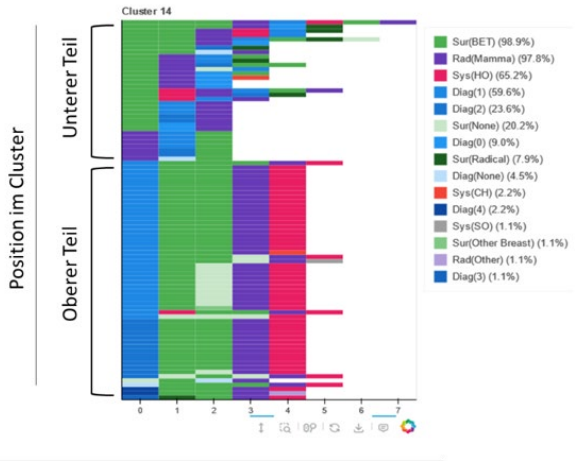


Abbildung 21: Behandlungssequenzansicht von Cluster 14 (TSH=1)

Beim direkt unten angrenzenden Beispiel Cluster 44, bzw. 4 und 148 war der gegenteilige Effekt zu sehen. Visuell scheint Cluster 44 (TSH 3) homogener, jedoch wurde es beim Verfahren mit der niedrigeren TSH auf zwei verschiedene Cluster aufgeteilt. Auch beim Betrachten der Behandlungssequenzansichten konnte kein auffälliger Unterschied zwischen Cluster 4 und Cluster 148 festgestellt werden (siehe Abbildung 22).

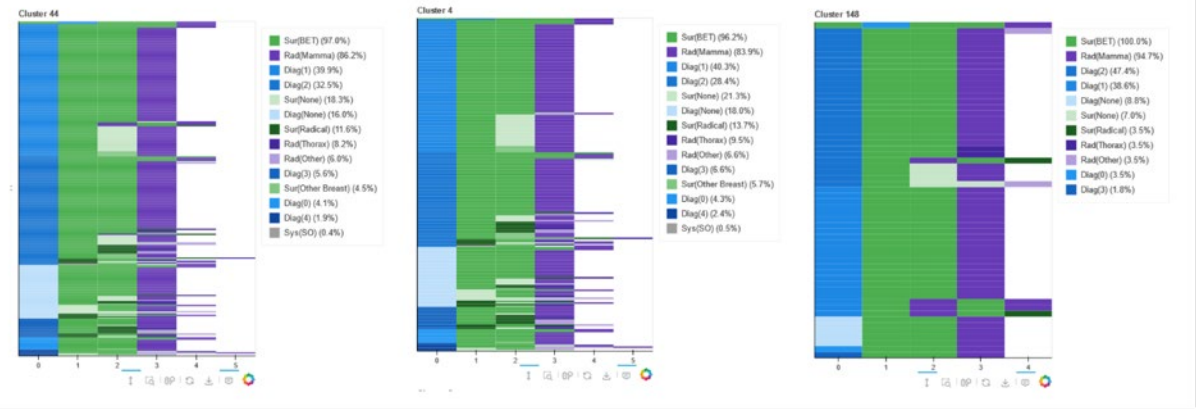


Abbildung 22: Behandlungssequenzansichten der Cluster 44 (TSH=3) und 4,148 (TSH=1)

Im Nachgang erst konnte geklärt werden, dass die Clustereinteilung an dieser Stelle vermutlich auf die Altersverteilung (c 4: 80% 50-69, c 148: <50 und >69) zurückzuführen ist.

Insgesamt wurde festgestellt: Sowohl plausible Clusteraufteilungen, bis auf Einzelfallebene sind zu finden, als auch nicht sofort ersichtliche Einteilungen der Behandlungsverläufe bei sehr hoher Clusterzahl.

3.3.4 Anwendungsmöglichkeiten

Es wurden mehrere Vorschläge geäußert, wie das Clustering der Behandlungsabläufe in der Zukunft in der Praxis angewendet werden könnte. Darunter war die Idee, dass man es zur

Überprüfung der Datenqualität nutzen könnte, in dem man sich auffällige oder auch unerklärliche Cluster genauer anschaut. Z. B. wurden Cluster mit Therapien vor Diagnose oder BET nach Mastektomie beobachtet und der Wunsch geäußert, diese näher zu untersuchen.

Eine andere Idee war es, bei den von der KAS regelmäßig veranstalten Qualitätskonferenzen mit den Meldern Clustering-Ergebnisse zu zeigen, zum einem, um das Interesse für das jeweilige Thema zu wecken und zum anderen, um diese noch weiteren Personengruppen (Ärzten, medizinischen Dokumentaren, etc.) zukommen zu lassen.

Auch die Möglichkeit QI-Ergebnisse als Cluster darzustellen wie in Abbildung 16 gezeigt stieß auf großes Interesse und könnte auf weitere Entitäten ausgeweitet werden, wobei zu beachten wäre, dass das bei unserem Modell auch viel Experten-Input benötigt.

Das Clustern als Vorauswahl zur Kohortenfindung zu nutzen war eine weitere Idee.

3.3.5 Anpassungsvorschläge

Abschließend wurden mögliche Anpassungswünsche an das Modell diskutiert, um es für den Einsatz in der Praxis zu optimieren. Einer davon war, die Daten vor dem Clustering stärker zu bereinigen, z.B. keine DCO- Fälle oder unplausible Therapiedaten zuzulassen. Auch kam der Wunsch den zeitlichen Abstand zwischen den einzelnen Ereignissen zu berücksichtigen. Das Unterscheiden von Erstlinien- und Zweitlinientherapie war eine weitere Idee, die einiges an Vorarbeit erfordern würde, ebenso wie der Gedanke, Multimorbidität als Variable zu berücksichtigen und das Clustern auf die Patientenebene auszuweiten. Auch den Tod oder die Zeit bis Tod könnte man beim Clustering miteinbeziehen oder wenigstens darstellen. Eine Umverteilung der Gewichtung zwischen Diagnose und Behandlung wurde ebenfalls überlegt, z.B. auch über die Einschränkung der Attribute bei der Diagnose.

3.4 Gesamtbeurteilung der inhaltlichen Plausibilität

Da bei vielen Clustern die Diagnosefaktoren (z.B. Alter und Grading) eine große Rolle spielen, könnte man bei einer Wiederholung des Verfahrens ins Betracht ziehen von vorneherein eine homogene Ausgangssituation (z.B. Tumore mit niedrigem oder hohem Risiko, etc.) herzustellen und die Diagnose nicht für das Clustering zu benutzen, um den Focus mehr auf die Behandlungsverläufe zu legen.

Andererseits zeigt aber auch das hier evaluierte Modell bereits sehr viele sich klar voneinander abgrenzenden Cluster mit konkreten Behandlungsverläufen, darunter sowohl bekannte und den Leitlinien entsprechende als auch eher ungewöhnliche oder welche, die auf mangelnde Datenqualität hinweisen.

All diese Behandlungseinteilungen sind durchaus sinnvoll und zeigen, dass die Grundidee, ein Clusterverfahren dazu zu benutzen, erfolgreich war.

3.5 Überlebenszeitanalysen basierend auf Clustern

3.5.1 Einleitung

Um die inhaltliche Validität der Cluster-Einteilungen weiter zu bewerten, wurde basierend auf den in den vorherigen Abschnitten beschriebenen Cluster-Verfahren eine Überlebenszeitanalyse durchgeführt. Ziel dieser Analysen war es bei Patientinnen mit ähnlichen prognostischen Tumoreigenschaften und ähnlichem Alter zu überprüfen, ob Cluster gebildet werden, die sich hinsichtlich des Überlebens unterscheiden, obwohl die Überlebenszeit selbst nicht in die Clusterbildung eingeflossen war. Eine solche Analyse hat das Potential im Rahmen von explorativen Analysen auffällig gute oder schlechte Krankheitsverläufe zu identifizieren und im Nachgang Behandlungspfade zu untersuchen, die mit diesen Verläufen assoziiert waren.

3.5.2 Methoden

Zu diesem Zweck wurde der Krebsregisterdatensatz zunächst auf Fälle aus dem Diagnosejahr 2019 mit invasivem Brustkrebs (ICD-10-Kode C50) und einem Diagnosealter zwischen 50 und 69 Jahren eingeschränkt. Zudem mussten die eingeschlossenen Patientinnen mindestens 56 Tage nach Diagnose überlebt haben. Innerhalb dieser Gruppe wurden zwei Subgruppen gebildet:

1) Prognostisch günstige Fälle (PG) mit UICC Stadien I oder II und
2) prognostisch ungünstige Fälle (PU) mit UICC Stadien III oder IV.

Diese Stratifizierung der Patientinnen wurde gewählt, um möglichst auszuschließen, dass sich die Überlebenszeiten zwischen den Clustern durch eine Häufung von Patientinnen mit besonders günstigen oder ungünstigen prognostischen Tumoreigenschaften unterscheiden. Im Idealfall sollten sich Überlebensunterschiede so überwiegend durch unterschiedliche Behandlungspfade ergeben.

Für diese Analyse wurde das bereits in den vorherigen Evaluationsschritten benutzte UMAP-Dimensionsreduktionsverfahren benutzt, um die multidimensionalen Abstände der Patientinnen auf zwei Dimensionen abzubilden. Die Auswahl der UMAP Parameter wurde basierend auf der visuellen Inspektion der resultierenden 2-D Abbildungen vorgenommen. Für die Modellauswahl des Clusterings wurden der DBSCAN-Algorithmus sowie das hierarchische Clustering mit verschiedenen Parameterkonfigurationen berechnet. Anschließend wurden die Ergebnisse auf die gewünschte Clusteranzahl von 15 – 30 Clustern eingeschränkt und das Modell mit dem niedrigsten Silhouette Score ausgewählt. Dieser Modellauswahlalgorithmus wurde für die PG- und PU-Gruppen separat durchgeführt. Die gewünschte Clusteranzahl von 15 – 30 wurde in Abstimmung mit den Fachexperten hinsichtlich einer noch interpretierbaren Clusteranzahl und einer, angesichts der zu erwartenden Menge an Fällen mit den gewünschten prognostischen Eigenschaften, ausreichend hohen Fallzahl pro Cluster festgelegt.

Für eine qualitative Beurteilung der Cluster wurden die jeweiligen medianen Behandlungsverläufe (Medoid des Clusters) benutzt, die zuvor grafisch aufbereitet wurden (s. Abbildung 23).

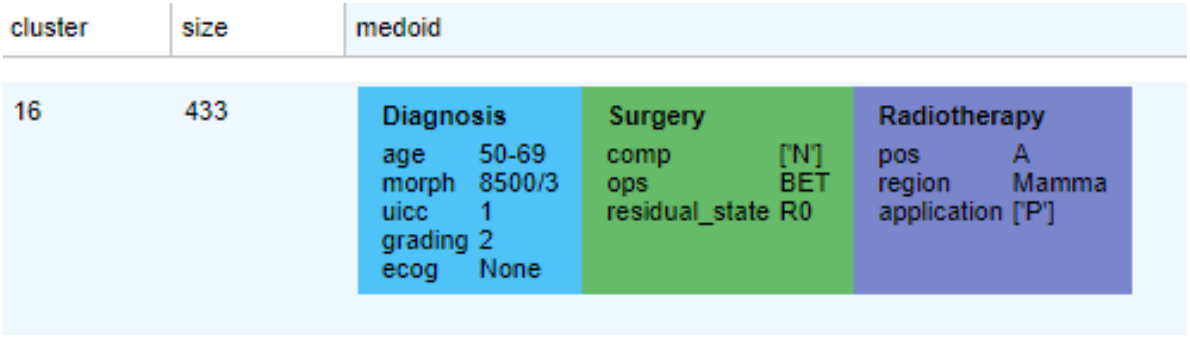


Abbildung 23: Grafische Aufbereitung des zentralen Falls (Medoid) am Beispiel von Cluster 16 der prognostisch günstigen Teilkohorte.

Die Überlebenszeit der Patientinnen in den ausgewählten Clustern wurde mit der Kaplan-Meier-Methode und einem Landmark-Ansatz [1] berechnet. Als Endpunkt wurde das Gesamtüberleben gewählt. Als Beginn der Überlebenszeit wurde das Diagnosedatum + 56 Tage (Landmark-Periode) definiert. Eine Rechtszensierung wurde an dem Tag des letzten Abgleichs aller Tumorfälle mit den Sterbefalldaten aus den Standesämtern vorgenommen (30.06.2023). Bedingt durch die Wahl des Diagnosejahres 2019 und das Zensierungsdatum wurde das 4-Jahres-Überleben für den Vergleich der Cluster ausgewählt. Cluster, in denen keine Sterbeereignisse vorkamen wurden von der Überlebenszeitanalyse ausgeschlossen.

Die Landmark-Methode wurde gewählt, da sowohl in der PG als auch in der PU Kohorte Patientinnen untersucht werden, für die im Krebsregister ausschließlich ein Diagnoseereignis und keine Informationen zu Therapien vorliegen. Bei Patientinnen, zu denen Therapieinformationen vorliegen, kann so das Phänomen von „Immortal Time“ – also einem Zeitintervall, in dem ein Versterben unmöglich war, weil die Person bis zu dem bekannten Beginn der Therapie überlebt haben musste – auftreten. Um die Gefahr eines Immortal-Time-Bias zu reduzieren

wurde ein Therapiebeginn nach spätestens 8 Wochen (56 Tage) nach Diagnosestellung angenommen [2] und es wurden nur Patientinnen in die Analyse eingeschlossen, die mindestens 8 Wochen nach Diagnosestellung überlebt hatten.

Auf der Kaplan-Meier-Analyse aufbauend wurde ein multiples Cox-Regressionsmodell benutzt um dem verbleibenden Confounding durch Unterschiede von histologischem Typ (lobuläres Karzinom vs. nicht-lobuläres-Karzinom), Diagnosealter, Grading und UICC-Stadium zu begegnen. Als Referenzkateogorie für die Clusterzugehörigkeit wurde das Cluster mit dem höchsten beobachteten 4-Jahresüberleben in der Kaplan-Meier-Analyse gewählt. Anschließend wurden mittels G-Computation Confounder-adjustierte Überlebenskurven berechnet, um die bereinigten Überlebenswahrscheinlichkeiten der Cluster grafisch darzustellen. Die Adjustierung erfolgte dabei mit den Parametern aus dem zuvor angepassten Cox-Modell.

Um die Behandlungsverläufe von Clustern mit besonders gutem bzw. schlechtem Überleben nach Adjustierung für mögliche Confounder zu illustrieren, wurden basierend auf den Cox Modellen jeweils das Cluster mit dem minimalen, dem medianen (im Falle einer geraden Clusteranzahl die beiden medianen Cluster) und dem maximalen Hazard-Ratio (HR) bezüglich ihrer Behandlungsverläufe genauer beschrieben.

Für die Datenaufbereitung und alle Analysen wurde das Softwarepaket R benutzt [3]. Die rohen und adjustierten Überlebenskurven wurden mit dem Paket adjustedCurves berechnet [4]. Für die Berechnung der Cox-Proportional-Hazards Modelle wurde das survival Paket benutzt [5].

3.5.3 Ergebnisse

Prognostisch günstige Fälle

Nach dem Ausschließen eines Clusters ohne Sterbeereignisse (Cluster 25, n = 138) bestand diese Subgruppe aus 4.830 Patientinnen mit einem Diagnosealter von durchschnittlich 59,6 Jahren (SD = 5,9 Jahre). Die Verteilung des UICC-Stadiums in der PG-Subgruppe war Stadium IA = 61,4%, Stadium IB = 2%, Stadium IIA = 27,1% und Stadium IIB = 9,5%. Das Grading verteilte sich auf 16,0% mit G1, 59,0% mit G2 und 25,0% mit G3. Der Anteil von Tumoren des lobulären histologischen Subtyps betrug 12.6%.

Für das Clustering der PG-Gruppe wurde ein hierarchisches Clustering mit einem distance threshold Parameterwert von 4,7 gewählt. Die Subgruppe wurde so in 28 Cluster mit einer durchschnittlichen Größe von n = 177 (SD = 140, Spannweite: 46 – 539) eingeteilt.

Die rohen 4-Jahres-Überlebenswahrscheinlichkeiten der einzelnen Cluster reichten von 87,1% [95%KI: 0,79 – 0,96] bis 0,99% [95% KI: 0,97 – 1,00] (s. Abbildung 24). Die Ergebnisse des Cox-Modells zeigten, dass auch nach Adjustierung für Diagnosealter, Stadium, Grading und Histologie weiterhin Unterschiede im Überleben zwischen den Clustern bestanden. Das HR für die Clusterzugehörigkeit lag zwischen 0,85 [95%KI: 0,05 – 13,72] in Cluster 10 und 12,38 [95%KI: 1,52 – 100,10] in Cluster 19 (s. Tabelle 4). Die beiden medianen Cluster waren Cluster 8 mit einem HR von 4,20 [95%KI: 0,53 – 33,37] und Cluster 15 mit einem HR von 4,79 [95%KI: 0,52 – 44,24]. Das prognostisch günstigste Cluster 10 bestand überwiegend aus Patientinnen, die eine brusterhaltende Operation mit adjuvanter Strahlentherapie erhalten hatten. Die beiden

bezüglich des HR medianen Cluster bestanden aus Patientinnen, für die ausschließlich eine brusterhaltende Operation mit R0-Status und ohne Vorliegen von Komplikationen gemeldet wurden (Cluster 8) beziehungsweise Patientinnen, zu denen dem LKR keine Therapieinformationen vorlagen (Cluster 15). Das prognostisch ungünstigste Cluster 19 beinhaltete Patientinnen, deren einzige vorliegende Therapie eine neoadjuvante Chemotherapie war.

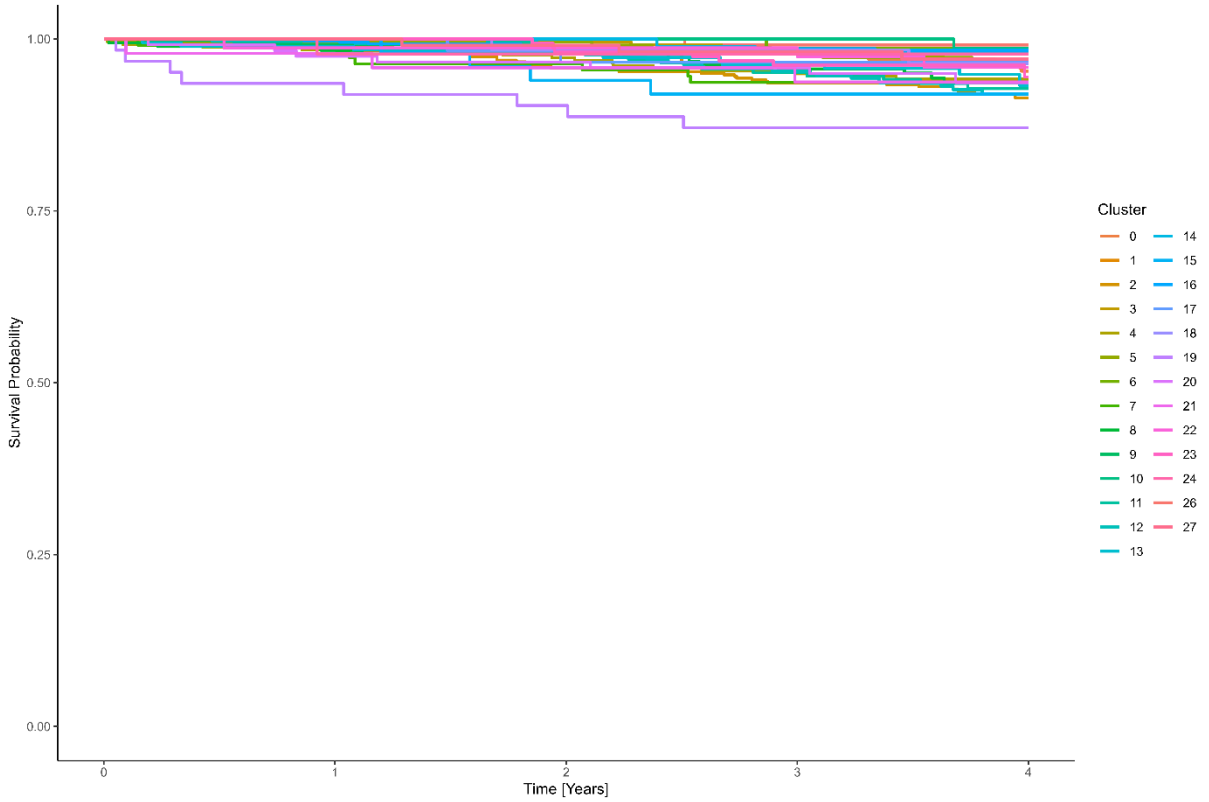


Abbildung 24: Nach Cluster stratifizierte, rohe Kaplan-Meier-Kurven der prognostisch günstigen Teilkohorte.

Tabelle 4: Ergebnisse des Cox-Proprtional-Hazards-Modells der prognostisch günstigen Brustkrebsfälle.

	Hazard ratio	Standardfehler	95%KI Untergenze	95%KI Obergenze
Cluster 0	3,92	1,03	0,52	29,34
Cluster 1	3,52	1,23	0,31	39,63
Cluster 2	6,88	1,02	0,93	51,19
Cluster 3	3,98	1,06	0,5	31,49
Cluster 4	6,52	1,04	0,86	49,64
Cluster 5	3,68	1,08	0,44	30,69
Cluster 6	1,35	1,43	0,08	22,04
Cluster 7	8,82	1,06	1,1	70,78
Cluster 8	4,2	1,06	0,53	33,37
Cluster 9	2,89	1,06	0,36	23,19

Cluster 10	0,85	1,42	0,05	13,72
Cluster 11	4,79	1,05	0,62	37,2
Cluster 12	6,55	1,03	0,86	49,67
Cluster 13	8,11	1,06	1,02	64,42
Cluster 14	2,67	1,23	0,24	29,51
Cluster 15	4,79	1,13	0,52	44,24
Cluster 16	2,11	1,08	0,26	17,4
Cluster 17	3,16	1,23	0,28	35,29
Cluster 18	3,38	1,23	0,31	37,43
Cluster 19	12,38	1,07	1,52	101
Cluster 20	5,29	1,08	0,64	44,02
Cluster 21	6,4	1,16	0,66	61,77
Cluster 22	5,47	1,16	0,56	53,02
Cluster 23	5,76	1,07	0,71	46,91
Cluster 24	6,16	1,23	0,56	68,31
Cluster 26 (Referenz)	1	-	-	-
Cluster 27	2,87	1,16	0,3	27,72
Nicht-lobuläres Ca. (Referenz)	1	-	-	-
Lobuläres Ca.	1,89	0,27	1,12	3,19
UICC IA (Referenz)	1	-	-	-
UICC 1B	1,26	0,59	0,4	4
UICC 2A	1,54	0,18	1,07	2,21
UICC 2B	2,79	0,21	1,86	4,18
Grading G1 (Referenz)	1	-	-	-
Grading G2	0,81	0,29	0,46	1,44
Grading G3	1,25	0,32	0,67	2,31
Diagnosealter	1,08	0,01	1,05	1,1

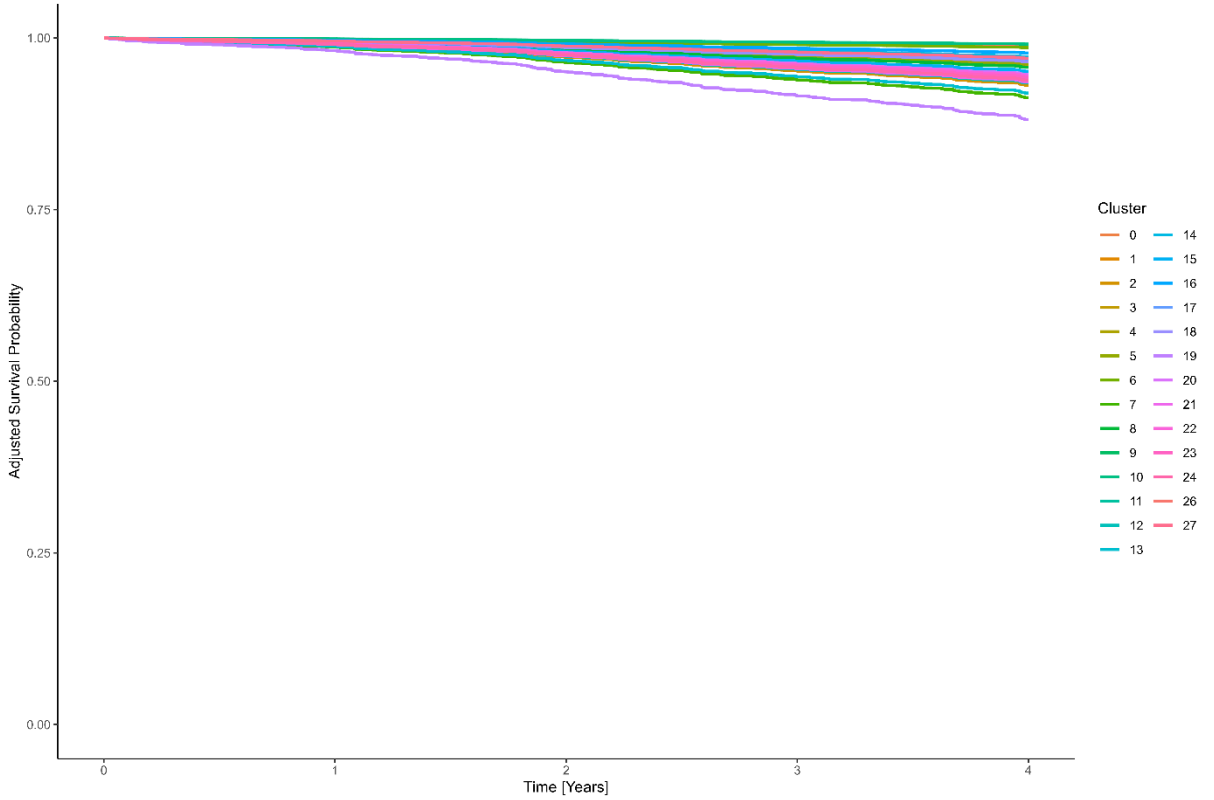


Abbildung 25: Nach Cluster stratifizierte, Kaplan-Meier-Kurven der prognostisch günstigen Teilkohorte. Adjustiert für Diagnosealter, Geschlecht, Grading, histologischem Subtyp und UICC-Stadium.

Prognostisch ungünstige Fälle

Auch in der PU-Gruppe wurde ein Cluster ohne Sterbeereignisse ausgeschlossen (Cluster 5, n = 6). Diese Gruppe bestand aus 267 Patientinnen mit einem Durchschnittsalter von 59,4 Jahren (SD =5,8 Jahre). Von ihnen waren bei Diagnose 20,0% in UICC-Stadium IIIA, 13,8% in Stadium IIIB, 14,9% in Stadium IIIC und 51,3% in Stadium IV. Bezüglich des Gradings waren 3,4% mit G1, 51,8% mit G2 und 44,9% mit G3 diagnostiziert worden.

Für die Subgruppe der PU-Fälle wurde das DBSCAN Modell mit den Parametern ε = 0,4 und minPts = 7 gewählt. So wurden 19 Cluster identifiziert, die eine durchschnittliche Größe von 32,8 Punkten (SD = 30,5, Spannweite 6 – 123) hatten.

Die 4-Jahres-Überlebenswahrscheinlichkeiten der Cluster reichten von 29,0% (95%CI: 16,7% - 50,39%) in Cluster 8 bis 90,9% (95%CI: 75,4% - 100,0%) in Cluster 10 (siehe Abbildung 25). Im Confounder-adjustierten Cox-Modell wurden Hazard-Ratios für die Clusterzugehörigkeit von 1 (Referenzkategorie der Cluster Variable) in Cluster 10 bis 11,00 [1,26 – 96,23] in Cluster 12 gefunden (s. Tabelle 5 + Abbildung 27). Das mediane Cluster bezüglich des adjustierten HR war Cluster 13 mit einem HR von 5,33 [95%KI: 0,59 – 48,41].

Charakteristisch für das Cluster 10 mit dem günstigsten Verlauf war eine brusterhaltende Operation mit R0-Status und anschließender adjuvanter Bestrahlung. Die typische Patientin im

Cluster mit dem medianen HR (Cluster 13) hatte eine Mastektomie mit R0-Status und anschließende adjuvante Strahlentherapie erhalten. Für Patientinnen in Cluster 12 (höchster HR) lagen dem LKR keine Behandlungsinformationen sondern lediglich eine Diagnose vor.

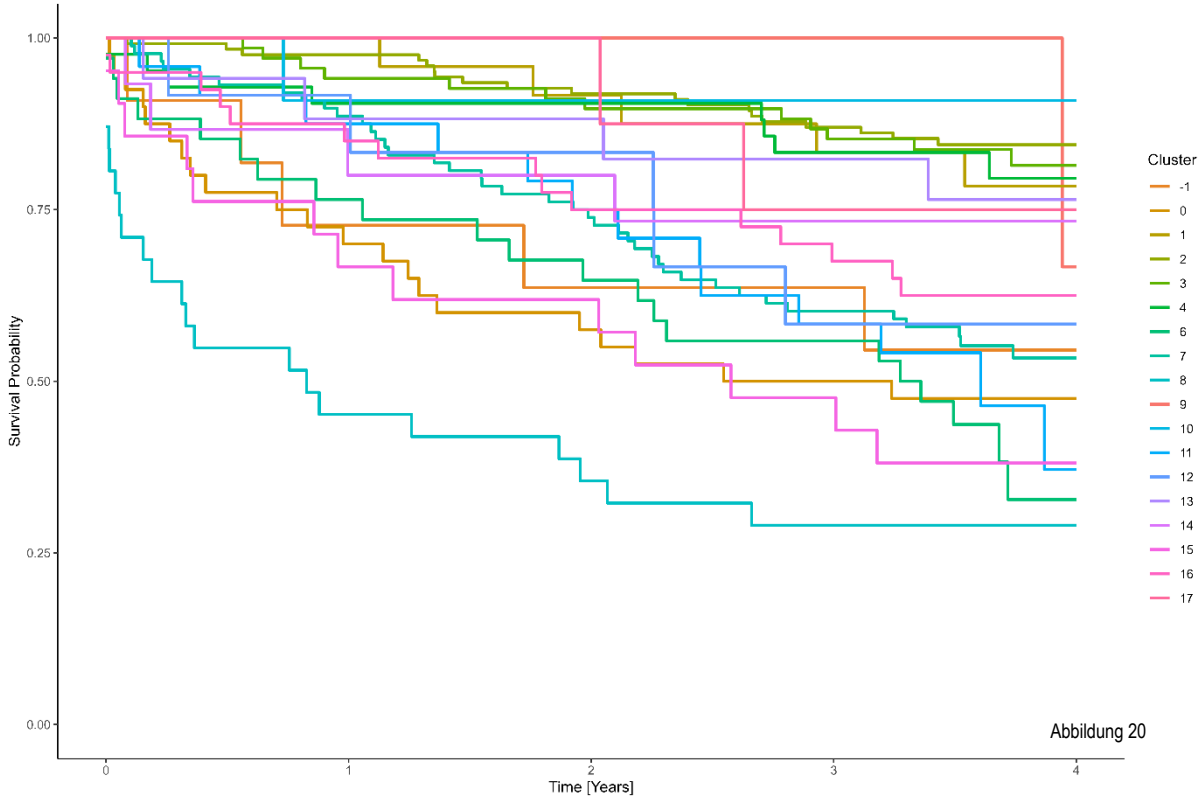


Abbildung 26: Nach Cluster stratifizierte, rohe Kaplan-Meier-Kurven der prognostisch ungünstigen Teilkohorte.

Tabelle 5: Ergebnisse des Cox-Proprtional-Hazards-Modells der prognostisch ungünstigen Brustkrebsfälle.

	Hazard ratio	Standardfehler	95%KI Untergenze	95%KI Obergenze
Cluster -1 (Ausreißer)	6,98	1,1	0,81	60,01
Cluster 0	9,53	1,03	1,26	72,18
Cluster 1	2,85	1,1	0,33	24,62
Cluster 2	2,63	1,03	0,35	19,7
Cluster 3	2,17	1,05	0,28	16,84
Cluster 4	2,62	1,06	0,33	21,01
Cluster 6	8,16	1,03	1,08	61,38
Cluster 7	5,32	1,02	0,72	39,02
Cluster 8	9,33	1,03	1,24	70,14
Cluster 9	1,5	1,42	0,09	24,22
Cluster 10 (Referenz)	1	-	-	-
Cluster 11	7,03	1,04	0,91	54,25
Cluster 12	11	1,11	1,26	96,23
Cluster 13	5,33	1,13	0,59	48,41

Cluster 14	3,54	1,13	0,39	32,4
Cluster 15	9,77	1,04	1,27	75,15
Cluster 16	6,79	1,04	0,89	51,64
Cluster 17	2,68	1,23	0,24	29,64
Nicht-lobuläres Ca. (Referenz)	1	-	-	-
Lobuläres Ca.	0,8	0,21	0,53	1,21
UICC 3A (Referenz)	1	-	-	-
UICC 3B	1,89	0,32	1	3,57
UICC 3C	1,22	0,35	0,61	2,43
UICC 4	2,69	0,29	1,52	4,76
Grading G1 (Referenz)	1	-	-	-
Grading G2	0,94	0,43	0,4	2,2
Grading G3	2,01	0,44	0,86	4,74
Diagnosealter	1,02	0,01	0,99	1,04

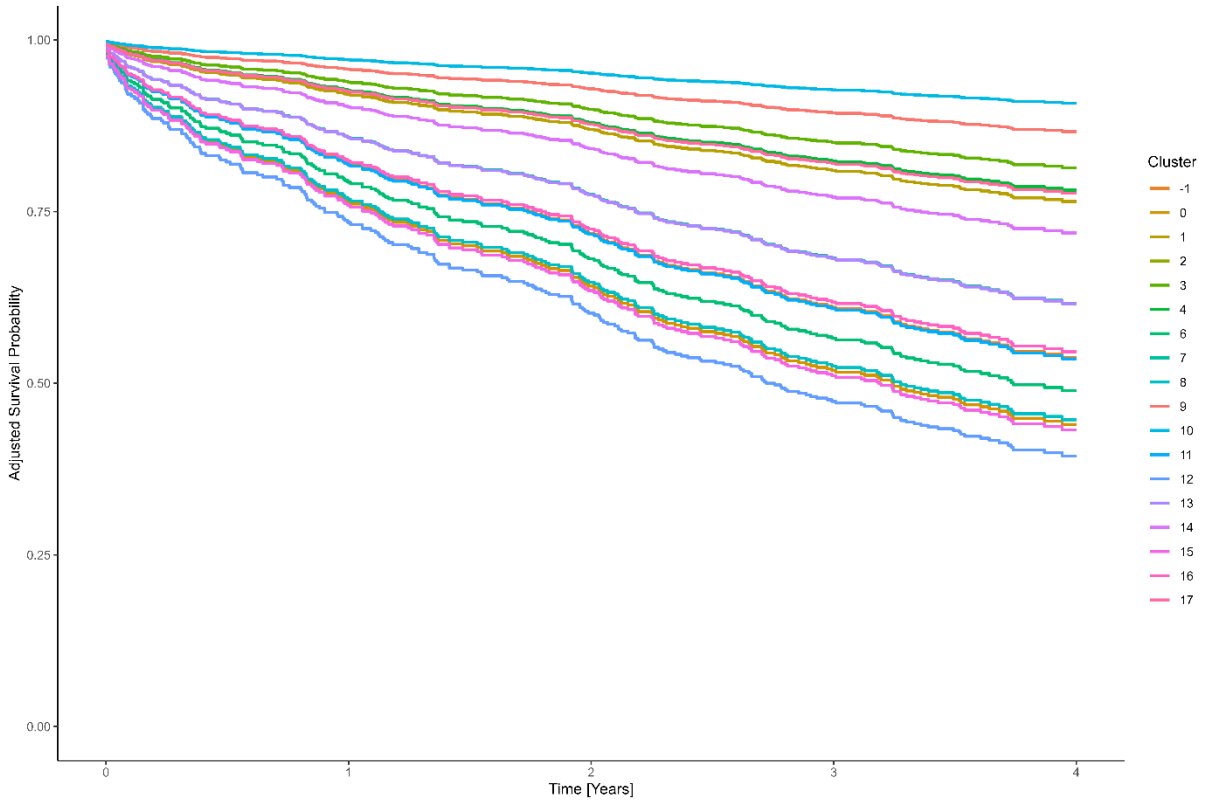


Abbildung 27: Nach Cluster stratifizierte, Kaplan-Meier-Kurven der prognostisch ungünstigen Teilkohorte. Adjustiert für Diagnosealter, Geschlecht, Grading, histologischem Subtyp und UICC-Stadium.

3.5.4 Diskussion

Die in den vorangegangenen Abschnitten entwickelten Methoden zum Clustering von Krebsregisterdaten anhand der Abstände von Behandlungsverläufen wurden hinsichtlich der Prognose der gefundenen Cluster angewendet. Dazu wurde der zugrundeliegende Datensatz auf Fälle eingeschränkt, die bereits ähnliche Tumoreigenschaften aufweisen, sodass sich etwaige Überlebensunterschiede vor allem durch Unterschiede in den Therapieverläufen erklären lassen sollten.

Die Vergleiche der rohen Überlebenskurven zeigten zum Teil deutliche Unterschiede im 4-Jahres-Überleben zwischen den Clustern, wobei diese deutlich stärker in der PU-Gruppe ausgeprägt waren. Diese Unterschiede lassen sich teilweise auf Unterschiede in den Basischarakteristika der Patientinnen zurückführen, da die PG- und PU-Gruppen nur anhand des UICC Stadiums definiert wurden. Nachdem im Cox-Proportional-Hazards-Modell für das genaue Stadium, Grading, histologischen Subtyp und Diagnosealter adjustiert wurde, reduzierten sich die Unterschiede zwischen den Clustern zwar, ein Clustereffekt war aber weiter zu beobachten.

In der PG-Gruppe war Cluster 10 auch nach Adjustierung das Cluster mit dem günstigsten Überleben. Patientinnen in diesem Cluster hatten eine brusterhaltende Operation und eine adjuvante Strahlentherapie erhalten, was der Empfehlung 4.86. in der S3-Leitlinie für das Mammakarzinom entspricht [6]. Eines der beiden Cluster, die den Median der HR ausmachten (Cluster 8), bestand aus Patientinnen, für die eine brusterhaltende Operation aber keine adjuvante Bestrahlung vorlag. Das deutlich schlechtere Überleben im Vergleich zu Cluster 10 kann so interpretiert werden, dass in diesem Cluster sowohl Patientinnen vertreten waren, die tatsächlich keine Strahlentherapie erhalten haben als auch solche, deren stattgehabte Strahlentherapie dem Register nicht gemeldet wurde. Das prognostisch ungünstigste Cluster 19 war durch eine neoadjuvante Chemotherapie als einzige bekannte Behandlung gekennzeichnet. Auch in diesem Fall kann nicht ausgeschlossen werden, dass weitere erfolgte Behandlungen dem Register nicht mitgeteilt wurden, andere Erklärungen für das Fehlen der weiteren Therapie in Kombination mit dem ungünstigen Verlauf könnten ein Versterben während der neoadjuvanten Therapie oder eine Verweigerung der Operation gewesen sein.

Die PU-Gruppe zeigte eine deutlich größere Spanne von Überlebensunterschieden zwischen den Clustern, die abgeschwächt auch nach der Confounder-Adjustierung bestehen blieben. Das prognostisch günstigste Cluster 10 bestand wie auch in der PG-Gruppe aus Fällen mit brusterhaltender Operation und adjuvanter Strahlentherapie. Das mediane Cluster bezüglich des HR dagegen bestand vorwiegend aus Frauen, die eine Mastektomie mit adjuvanter Strahlentherapie erhalten hatten. Es ist plausibel, dass das Überleben im Vergleich zu Cluster 10 schlechter war, da z.B. ein lokal weiter ausgedehnter Tumor vorlag, sodass keine brusterhaltende OP möglich war. Für Patientinnen im prognostisch ungünstigsten Cluster 13 lag ausschließlich eine Diagnose vor, sodass nur Mutmaßungen über das Fehlen der Therapien im Krebsregister oder eine nicht-stattgefundene Therapie möglich wären.

Es ist wichtig zu betonen, dass alleine aus den Ergebnissen dieser Auswertung keine Rückschlüsse über die Effektivität der beschriebenen Behandlungsstrategien möglich sind. Es ist bekannt, dass Therapieinformationen dem Krebsregister nur unvollständig vorliegen. Es ist zudem wahrscheinlich, dass die Clustereffekte von weiterem unbeobachteten Confounding betroffen sind. Die Landeskrebsregister verfügen zum Beispiel nicht über Informationen zu

Komorbiditäten, die sowohl die Behandlungsentscheidung als auch das Gesamtüberleben beeinflussen können. Auch der Hormonrezeptor- bzw. HER2/NEU-Status, die besonders beim Mammakarzinom eine wichtige prognostische und therapeutische Rolle spielen, konnten in dieser Auswertung nicht berücksichtigt werden. Eine weitere Herausforderung stellte die relativ geringe Fallzahl in der PU Gruppe dar, die zur Folge hatte, dass die HR der einzelnen Cluster nicht mit einer zufriedenstellenden Präzision geschätzt werden konnten. Eine Grundsätzliche Limitation ist die Methodik der Abstandsberechnungen und der Auswahl des Clustering-Verfahrens, die viele subjektive Entscheidungen des Auswerters erfordern und nur schwer reproduzierbar sind.

Zusammenfassend zeigt sich dennoch, dass das Clustering - auch ohne Berücksichtigung des Überlebens bei der Berechnung der Abstandsmaße der Fälle - prognostisch deutlich abgrenzbare Gruppen von Therapieverläufen findet, die sich nicht allein durch unterschiedliche Verteilungen der Baseline-Charakteristika der Patientinnen erklären lassen. Der Ansatz hat das Potential die Krebsregisterdaten explorativ auszuwerten und Auswerter auf unerwartete Zusammenhänge von Überlebenswahrscheinlichkeit und Therapieverlauf aufmerksam zu machen. Diese Ergebnisse können unter anderem dabei helfen, neue Forschungshypothesen zu generieren und mangelnde Daten- und Versorgungsqualität aufzudecken.

4. Diskussion der Evaluationsergebnisse

In der Entwicklung der des DMW zeigte sich, dass Diagnosefaktoren wie Alter und Grading eine zentrale Rolle spielen. Für die Weiterentwicklung des Tools könnte es sinnvoll sein, die Ausgangsdaten homogener zu gestalten, um den Fokus stärker auf Behandlungsverläufe zu legen, ohne Diagnosen als Clustering-Kriterium zu verwenden. Das aktuelle Modell zeigt bereits klar abgrenzbare Cluster mit typischen und ungewöhnlichen Behandlungsverläufen, was die Effektivität des Verfahrens bestätigt. Clustering offenbart unerwartete Zusammenhänge zwischen Überlebenswahrscheinlichkeit und Therapieverlauf und bietet Potenzial zur Hypothesengenerierung und Qualitätskontrolle, obwohl mögliche Verzerrungen durch unbekannte Faktoren wie Komorbiditäten zu beachten sind.

Die Limitationen des Tools umfassen das Fehlen einer Benutzeroberfläche für DMW und die Abhängigkeit von Programmierkenntnissen. Zudem erschweren die Datenqualität (insbesondere fehlende Werte), die erforderliche manuelle Definition der Abstandsmaße sowie die hohe Zahl an manuell zu wählenden Parametern die Reproduzierbarkeit der Ergebnisse. Analysen können nur auf speziell aufbereiteten Datensätzen durchgeführt werden, die an die jeweilige Struktur der Krebsregister angepasst werden müssen.

Zusammenfassend lässt sich sagen, dass das DMW die im SePaMiM Projekt adressierte Forschungsfrage – Behandlungsverläufe aus Krebsregisterdaten für die Versorgungsforschung nutzbar zu machen – gut nutzen. Durch geeignete Datenmodelle und Suchverfahren lassen sich spezifische Kohorten bilden und analysieren. Im Gegensatz zum PMAW Tool bzw. herkömmlichen Abfragesprachen ermöglicht das DMW die Identifikation von Patientenkohorten anhand der Ähnlichkeit von Behandlungsverläufen und ohne eine genaue Suchanfrage des Nutzers. Diese Eigenschaft lässt sich auf die im Projekt definierten sekundären Forschungsfragen nach der Etablierung von Standards für Routineauswertungen, die Erkennung alternativer Behandlungsverläufe sowie der Erkennung von Datenqualitätsproblemen übertragen. Die Ergebnisse der Überlebenszeitanalyse deuten zudem darauf hin, dass sich die so gefundenen Kohorten sinnvoll mit Routineanalysen, wie sie in Krebsregistern zu Anwendung kommen, ausgewertet werden können.

Literaturverzeichnis

1. Dafni U. Landmark Analysis at the 25-Year Landmark Point. Circulation: Cardiovascular Quality and Outcomes. 2011;4(3):363-371. doi:10.1161/CIRCOUTCOMES.110.957951

2. Rubio IT, Marotti L, Biganzoli L, et al. EUSOMA quality indicators for non-metastatic breast cancer: An update. European Journal of Cancer. 2024;198. doi:10.1016/j.ejca.2023.113500

3. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2023. https://www.R-project.org/

4. Denz R, Klaaßen-Mielke R, Timmesfeld N. A comparison of different methods to adjust survival curves for confounders. Statistics in Medicine. 2023;42(10):1461-1479.

5. Therneau TM. A Package for Survival Analysis in R.; 2020. https://CRAN.R-project.org/package=survival

6. Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF): S3-Leitlinie Früherkennung, Diagnose, Therapie und Nachsorge des Mammakarzinoms, Version 4.4, 2021, AWMF Registernummer: 032-045OL. Published online 2021:467.

Bochum, 10.01.2025

Florian Oesterling
Verantwortlicher Methodiker
LKR NRW, Bochum

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

Gefördert vom G-BA Innovationsfond
FKZ: 01VSF20018

PMAW CQL UND SQL VERGLEICHS ANFRAGEN DER EVALUATION

11. MÄRZ 2025

CHRISTIAN LÜPKES, KOLJA BLOHM
OFFIS – Institut für Informatik
Escherweg 22, 26121 Oldenburg

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

1 Präambel

In diesem Dokument werden für die verschiedenen, in SePaMiM betrachteten Bereiche der Evaluation jeweils ein Fallbeispiel erläutert.
Die insgesamt neun unterschiedlichen Fragestellungen aus dem LKR.NRW und der KLast wurden mit CQL implementiert und mittels PMAW ausgeführt. Diese PMAW Ergebnisse wurden dann mit den Ergebnissen aus bereits in den Krebsregistern etablierten Methoden (in der Regel SQL Anfragen) verglichen.
Sämtliche CQL Anfragen für alle Fallbeispiele und die SQL-Anfragen für den Goldstandard sind in der Softwaresammlung enthalten. Das Tooling der KLast für den Goldstandard der Routineauswertungen ist proprietär und deshalb nicht in der Softwaresammlung enthalten.
Die Beschreibungstexte sind dem Ergebnisbericht – Abschnitt 3.1.3. Evaluation und dem PMAW Evaluationsbericht entnommen. In letzterem findet sich zu allen Fallbeispielen eine ausführliche Beschreibung.

2 Qualitätsindikatoren – 1a

In diesem Abschnitt wird exemplarisch für das Fallbeispiel 1a der Evaluation zum einen die Definition durch die Plattform § 65 c für die Krebsregister als auch deren konkrete Umsetzung in eine CQL Abfrage mittels PMAW sowie als Referenzimplementierung / Goldstandard im Projekt SePaMiM die im LKR.NRW genutzte Definition abgedruckt.

Qualitätsindikatoren für Leitlinien sind Kennzahlen, die von der Arbeitsgruppe QI der Plattform § 65c von S3 Leitlinen abgeleitet werden, um die leitliniengerechte Behandlung von Krebspatient*innen zu bewerten. QI 8 stammt aus der S3- Leitlinie Mammakarzinom Version 4.4, Juni 2021. Nach brusterhaltender Operation wegen eines invasiven Karzinoms soll eine Bestrahlung der betroffenen Brust durchgeführt werden. Das Qualitätsziel ist eine adäquate Rate an Bestrahlungen nach BET bei Patienten mit Ersterkrankung invasives Mammakarzinom.

2.1 QI 8 Definition der Plattform § 65c

Abrufbar unter (https://plattform65c.atlassian.net/wiki/spaces/LLQI/pages/100600499/QI+8+-+Durchgef+hrte+Strahlentherapie+nach+BET)
Stand der Definition 22. Juni 2023. Word-Export von der Plattform § 65c Webseite durchgeführt am 08.01.2025.

QI 8 - Durchgeführte Strahlentherapie nach BET

Qualitätsziel LL	Adäquate Rate an Bestrahlungen nach BET bei Pat. mit Ersterkrankung invasives Mammakarzinom
Zähler LL	Pat. mit invasivem Karzinom und BET, die eine Radiatio der Brust erhalten haben
Nenner LL	Pat. mit Primärerkrankung invasives Mammakarzinom und BET
AG Operationalisierung LL-QI:	

Beurteilung	oBDS 2014: oBDS 2021: Berechenbar
Stand Operationalisierung	
Zähler	Therapiemeldung Strahlentherapie **MIT** (Stellung zur OP = adjuvant **ODER** Datum OP < Datum Start Strahlentherapie < Datum OP + 8 Monate) **MIT** Zielgebiet = „Mamma als Teilbrust“ (3.2.) **ODER** „Mamma als Ganzbrust“ (3.1.)
Nenner	ICD-10: C50* Primärerkrankung (kein Rezidiv) **UND** Histologie zu C50*: 8010/3-8570/3 und ungleich M1 **UND** Therapiemeldung OP; brusterhaltende OP => OPS: 5-870* **OHNE** Mastektomie OPS: 5-877*, 5-872*, 5-874*

Ergänzende Anmerkungen

- Allgemeine methodische Festlegungen beachten
- Ergänzende Anmerkungen AG QI §65c (zugangsbeschränkt)

2.2 CQL Abfrage PMAW

Im Folgenden ist das im Rahmen der Evaluation für PMAW in CQL geschriebene Abfrageskript wiedergegeben:

```
using SepamimModel

context Tumor
```

```
// Mammakarzinom QI 8 - Durchgeführte Strahlentherapie nach BET
//
// Nenner:
// - ICD-10 Diagnosecode(s): C50.0 – C50.9
// - Bei Diagnose nicht metastasiert: TNM M != 1
// - Histologie zu C50*: 8010/3-8570/3
// - Mindestens eine OP mit OPS Code 5-Steller = 5-870 vorhanden (Definition BET)
// - Keine OP mit OPS Codes für Mastektomie (5-872, 5-874, 5-877)
//
// Zähler:
// - Alles aus dem Nenner
// - Mindestens eine Meldung zu Bestrahlung, Zielgebiet: 3.1 (Mamma als Ganzbrust) oder 3.2 (Mamma als
Teil-brust)
// - Stellung der Bestrahlung zur BET = adjuvant ODER
// - Beginndatum der Bestrahlung ≥ OP-Datum UND Beginndatum ≤ OP-Datum + 8 Monate


// - ICD-10 Diagnosecode(s): C50.0 – C50.9
// - Bei Diagnose nicht metastasiert: TNM M != 1
// - Histologie zu C50*: 8010/3-8570/3
define "Initiale Population":
  StartsWith(Tumor.icdCode, 'C50')
  and not StartsWith(Coalesce(Tumor.ptnmm, Tumor.ctnmm, ''), '1')
  and Tumor.morphologyCodeIcdO in
  { '8500/3', '8022/3', '8035/3', '8520/3', '8211/3', '8201/3',
   '8480/3', '8510/3', '8513/3', '8507/3', '8575/3', '8570/3',
   '8572/3', '8070/3', '8032/3', '8571/3', '8575/3', '8982/3',
   '8246/3', '8041/3', '8574/3', '8502/3', '8503/3', '8550/3',
   '8430/3', '8525/3', '8290/3', '8314/3', '8315/3', '8410/3',
   '8983/3', '8200/3', '8504/3', '8509/3', '9020/3', '8530/3',
   '8000/3', '8001/3', '8005/3', '8010/3', '8012/3', '8013/3',
   '8020/3', '8021/3', '8033/3', '8050/3', '8071/3', '8074/3',
   '8076/3', '8140/3', '8141/3', '8190/3', '8230/3', '8260/3',
   '8265/3', '8249/3', '8310/3', '8401/3', '8470/3', '8481/3',
   '8490/3', '8508/3', '8512/3', '8521/3', '8522/3', '8523/3',
   '8524/3', '8541/3', '8543/3', '8560/3', '8570/3', '8573/3',
   '8815/3', '8830/3', '8854/3' }
  and Tumor.dateOfDiagnosis after or on @2016-04-01


// - Mindestens eine OP mit OPS Code 5-Steller = 5-870 vorhanden (Definition BET)
define "Brusterhaltende Therapien":
 [Surgery] bet
  where exists bet.ops opsCode
   where StartsWith(opsCode, '5-870')

// - Keine OP mit OPS Codes für Mastektomie (5-872, 5-874, 5-877)
define "Mastektomien":
 [Surgery] surgery
```

```
  where exists surgery.ops opscode
   where StartsWith(opscode, '5-872')
    or StartsWith(opscode, '5-874')
    or StartsWith(opscode, '5-877')

define Nenner:
 "Initiale Population"
  and exists "Brusterhaltende Therapien"
  and not exists "Mastektomien"

// - Mindestens eine Meldung zu Bestrahlung, Zielgebiet: 3.1 (Mamma als Ganzbrust) oder 3.2 (Mamma
als Teil-brust)
define "Relevante Bestrahlungen":
 [Radiotherapy] r
  where exists r.radiations radiation
   where StartsWith(radiation.targetRegion, '3.1')
    or StartsWith(radiation.targetRegion, '3.2')

// - Stellung der Bestrahlung zur BET = adjuvant ODER
// - Beginndatum der Bestrahlung ≥ OP-Datum UND Beginndatum ≤ OP-Datum + 8 Monate
define "Bestrahlungen nach BETs":
 from
  "Brusterhaltende Therapien" bet,
  "Relevante Bestrahlungen" rb
  where ( rb.surgeryPosition = 'A' )
   or start of rb.period between bet.date + 1 day and bet.date + 8 month

define Zaehler:
 "Nenner"
  and exists "Bestrahlungen nach BETs"
```

2.3 SQL Anfrage LKR.NRW

Im Folgenden ist das im Rahmen der Evaluation für PMAW im LKR.NRW verwendete Abfrageskript wiedergegeben, deren Ergebnis als Goldstandard gilt. Das Skript gibt die Filterkriterien wieder:

```
with
elem as(
  select et.patientid, et.tumorid, es.surgerydate_date, string_agg(ops,'||') as ops
  from element_tumorsynopsis et
  join event_surgery es using (patientid,tumorid)
  join esurgery_ops eo on eo.esurgery_id = es.id
  where et.icdcode like 'C50%'
  and et.morphologycodeicdo in
('8500/3','8022/3','8035/3','8520/3','8211/3','8201/3','8480/3','8510/3','8513/3',
  '8507/3','8575/3','8570/3','8572/3','8070/3','8032/3','8571/3','8575/3','8982/3','8246/3','8041/3','8574
/3','8502/3',
```

```
  '8503/3','8550/3','8430/3','8525/3','8290/3','8314/3','8315/3','8410/3','8983/3','8200/3','8504/3','8509
/3','9020/3',
  '8530/3','8000/3','8001/3','8005/3','8010/3','8012/3','8013/3','8020/3','8021/3','8033/3','8050/3','8071
/3','8074/3',
  '8076/3','8140/3','8141/3','8190/3','8230/3','8260/3','8265/3','8249/3','8310/3','8401/3','8470/3','8481
/3','8490/3',
  '8508/3','8512/3','8521/3','8522/3','8523/3','8524/3','8541/3','8543/3','8560/3','8570/3','8573/3','8815
/3','8830/3','8854/3')
  and left(coalesce(ptnmm,ctnmm),1) is distinct from '1'
  and dateofdiagnosis_date >= to_date('01.04.16','dd.mm.yy')
  group by 1,2,3
),
allsur as (
  select patientid, tumorid, string_agg(eo.ops,'||') as allops
  from elem
  join event_surgery es using (patientid,tumorid)
  join esurgery_ops eo on eo.esurgery_id = es.id
  group by 1,2
),
qi as (
  select elem.patientid, elem.tumorid,ops,
      case when (er.radiotherapysurgeryposition = 'A' or er.radiotherapystartdate_date between
surgerydate_date + 1 and surgerydate_date + interval '8 month')
          and ra.radiationtargetregion like any(array['3.1.%','3.2.%']) then 1
      else 0 end as qi7
  from elem
  join allsur using (patientid,tumorid)
  left join event_radiotherapy er on elem.patientid=er.patientid and elem.tumorid = er.tumorid
  left join event_radiation ra on ra.radiotherapy_id=er.id
  where ops like '%5-870%'
  and allops not like all(array['%5-877%','%5-872%','%5-874%'])
)

select 'NRW' as parentid,
    count(distinct patientid || '_' || tumorid) filter(where QI7=1) as zaehler, --Zähler
    count(distinct patientid || '_' || tumorid) as nenner --Nenner (=Gesamtzahl)
from qi ee
```

2.4 Ergebnisse

Folgende Ergebnisse erbrachten die Anfragen im Rahmen der Evaluation. Es gab keine Abweichungen zwischen den Ergebnissen der PMAW CQL Anfragen und dem SQL des LKR.NRW.

Anfrage	Zähler	Nenner
CQL	19078	41134
SQL NRW	19078	41134

3 Plausibiltätsprüfung – 2a

In diesem Abschnitt wird exemplarisch aus dem Bereich „Plausibilitätsprüfungen“ der Evaluation für das Fallbeispiel 2a die konkrete Umsetzung in eine CQL Abfrage mittels PMAW sowie als Referenzimplementierung / Goldstandard im Projekt SePaMiM die im LKR.NRW genutzte Abfrage abgedruckt.

Im Bereich *Plausibilitätsprüfung* wurden drei Fragestellungen definiert: die Suche nach klinischen Ereignissen nach dem Tod eines Patienten, die Suche nach Patient*innen die im Laufe Ihrer Behandlung von ein einer palliativen zu einer kurativen Therapieintention gewechselt sind und die Suche nach Patient*innen bei denen ein Rückgang der Tumorgröße gemessen wurde, ohne dass zwischen beiden Größenmessungen eine Therapie erfolgt ist. Da eine aussagekräftige Analyse der klinischen Fragestellungen nur gelingen kann, wenn die zugrundeliegenden Daten eine hohe Qualität aufweisen stehen bei diesen Fallbeispielen keine klinischen Fragestellungen, sondern die Prüfung der Datenqualität im Vordergrund.

3.1 Erläuterung LZG.NRW zu Fallbeispiel 2a - Klinische Ereignisse nach dem Tod des Patienten

Das Todesdatum kann aus mehreren Quellen stammen, es kann direkt von einer Einrichtung gemeldet werden oder durch regelmäßige Abgleiche mit den Einwohnermeldeämtern in unsere Datenbank aufgenommen werden. Sollten mehrere Therapieereignisse nach dem Tod des Patienten von verschiedenen Meldern gefunden werden, ist zu vermuten, dass ein Datum in der Datenbank falsch ist. Neben der Möglichkeit einer falschen Meldung könnte es auch ein Hinweis auf einen Fehler bei der Zuordnung von Todesmeldungen mit Krebsregisterfällen sein.

Operationalisierung der Filterkriterien

- Todesdatum < Op-Datum
- Todesdatum < Strahlentherapie Startdatum ODER
  Todesdatum < Strahlentherapie Enddatum ODER
  Todesdatum < Bestrahlung Startdatum ODER
  Todesdatum < Bestrahlung Enddatum ODER
- Todesdatum < Systemische Therapie Startdatum ODER
  Todesdatum < Systemische Therapie Enddatum ODER

Hinweis: Das Enddatum von Therapien wird oft nicht gemeldet, daher muss auch das Startdatum in die Abfrage mit ein bezogen werden. Zu jeder Strahlentherapie werden auch einzelne Bestrahlungsdaten gelmeldet.

3.2 CQL Abfrage PMAW

Im Folgenden ist das im Rahmen der Evaluation für PMAW in CQL geschriebene Abfrageskript wiedergegeben:

```
using SepamimModel
```

```
context Tumor

define "Initiale Population":
 true

define "Alle OP Datumswerte":
 from [Surgery] surgery return {Datum:surgery.date}

define "Hat Strahlentherapiedatum nach Todesdatum":
 exists [Radiotherapy] rt where
  rt.period ends after Tumor.death.date
  or exists rt.radiations r where
  r.startDate after Tumor.death.date
  or r.endDate after Tumor.death.date

define "Alle Strahlentherapie Datumswerte":
 (from [Radiotherapy] rt return {Datum: start of rt.period})
 union
 (from [Radiotherapy] rt return {Datum: end of rt.period})

define "Alle Systemtherapie Datumswerte":
 (from [SystemicTherapy] st return {Datum: start of st.period})
 union
 (from [SystemicTherapy] st return {Datum: end of st.period})

define "OPs nach Todesdatum":
 "Initiale Population"
  and exists "Alle OP Datumswerte" op where op.Datum after Tumor.death.date

define "Strahlentherapien nach Todesdatum":
 "Initiale Population"
  and "Hat Strahlentherapiedatum nach Todesdatum"

define "Systemtherapien nach Todesdatum":
 "Initiale Population"
  and exists "Alle Systemtherapie Datumswerte" st where st.Datum after Tumor.death.date
```

3.3 SQL Anfrage LKR.NRW

Im Folgenden ist das im Rahmen der Evaluation für PMAW im LKR.NRW verwendete Abfrageskript wiedergegeben, deren Ergebnis als Goldstandard gilt. Das Skript gibt die Filterkriterien wieder:

```
with
ac as (
  select patientid, tumorid, ed.dateofdeath_date
  from element_tumorsynopsis et
  join element_death ed using(patientid)
),
```

```
sys as (
  select patientid, tumorid, min(es.systemictherapystartdate_date)
  from ac
  join event_systherapy es using(patientid, tumorid)
  where es.systemictherapyenddate_date > ac.dateofdeath_date or es.systemictherapystartdate_date >
dateofdeath_date
  group by 1,2
),
sur as (
  select patientid, tumorid, min(es2.surgerydate_date)
  from ac
  join event_surgery es2 using(patientid, tumorid)
  where surgerydate_date > ac.dateofdeath_date
  group by 1,2
),
rad as (
  select patientid, tumorid, min(er.radiationstartdate_date)
  from ac
  join event_radiation er using(patientid, tumorid)
  join event_radiotherapy er3 using(patientid, tumorid)
  where radiationenddate_date > ac.dateofdeath_date or radiationstartdate_date > dateofdeath_date
  or er3.radiotherapyenddate_date > dateofdeath_date or er3.radiotherapystartdate_date  >
dateofdeath_date
  group by 1,2
)
select 'Systemische Therapie' as "Therapie", count(*)
from sys
union
select 'OP' as "Therapie", count(*)
from sur
union
select 'Radiatio' as "Therapie", count(*)
from rad
order by "Therapie";
```

3.4 Ergebnisse

Folgende Ergebnisse erbrachten die Anfragen im Rahmen der Evaluation. Es gab keine Abweichungen zwischen den Ergebnissen der PMAW CQL Anfrage und dem SQL des LKR.NRW.

Anfrage	OP	Radiotherapie	Systemtherapie
CQL	176	197	1611
SQL NRW	176	197	1611

4 Externe Datennutzung – 3a

In diesem Abschnitt wird exemplarisch aus dem Bereich „Externe Datennutzung“ der Evaluation für das Fallbeispiel 3a die konkrete Umsetzung in eine CQL Abfrage mittels PMAW sowie als Referenzimplementierung / Goldstandard im Projekt SePaMiM die in der KLast genutzte Abfrage abgedruckt.

Auf Anfrage können klinische Krebsregister Daten für Forschungszwecke an externe Forscher*innen bereitstellen. Im Bereich *Externe Datennutzung* wurden beispielhaft zwei Anträge für anonyme Einzelfalldaten, die an die KLast gestellt wurden, mit CQL umgesetzt. Hierbei wurden Daten für zwei Studien von einer Universitätsmedizin angefragt, um den Einfluss des Therapiebeginns bei fortgeschrittenem (UICC IV) kolorektalem, bzw. Mammakarzinom auf das Überleben zu untersuchen.

4.1 Erläuterung zum Fallbeispiel 3a – Brustkrebs

Bei der Datenanfrage an die KLast durch eine Universitätsmedizin ging es um die Durchführung einer Studie mit dem Ziel den Einfluss des Therapiezeitpunkts auf das Überleben bei Patienten mit fortgeschrittenem Mammakarzinom (UICC IV) zu untersuchen.

Es wurde nach ICD-Code des Tumors und dem Vorhandensein einer palliativen systematischen Therapie, sowie einigen typischen weiteren Merkmalen gefiltert, darunter Diagnosezeitraum und Geschlecht des Patienten.

4.2 CQL Abfrage PMAW

Im Folgenden ist das im Rahmen der Evaluation für PMAW in CQL geschriebene Abfrageskript wiedergegeben:

```
// Externe Anfrage zu Pilotstudie, um den Einfluss des Therapiezeitpunkts
// bei fortgeschrittenem Mammakarzinom zu untersuchen


using SepamimModel

context Tumor

define "Initial Population":
 StartsWith(Tumor.icdCode, 'C50')
  and PositionOf('M1', Tumor.tnmm) >= 0
  and Tumor.sex in { 'M', 'W' }
  and not EndsWith(Tumor.morphologyCodeIcdO, '/6')
  and Tumor.dateOfDiagnosis after @2018-07-01
  and exists ( [SystemicTherapy] s
    where s.intention = 'P'
  )
```

```
define Output:
 {
  tumorId: Tumor.tumorId,
  patientId: Tumor.patientId,
  age: Tumor.ageAtDiagnosis,
  sex: Tumor.sex,
  icd: Substring(Tumor.icdCode, 0, 3),
  yearOfDiagnosis: year from Tumor.dateOfDiagnosis,
  morphology: Tumor.morphologyCodeIcdO,
  side: Tumor.laterality,
  grading: Tumor.grading,
  generalPerformanceState: Tumor.generalPerformanceState,
  deadDaysAfterDiagnosis: DaysSinceDiagnosis(Tumor.death.date),
  tnmT: Tumor.tnmt,
  tnmN: Tumor.tnmn,
  tnmM: Tumor.tnmm,
  metastasis: metastasisOutput,
  surgery: surgeryOutput,
  radiotherapy: radioOutput,
  systemictherapy: systemicOutput
 }

define function DaysSinceDiagnosis(date Date):
 duration in days between Tumor.dateOfDiagnosis and date

define metastasisOutput:
 [Metastasis] m
  return all {
   "TumorId": Tumor.tumorId,
   "FM_Lokalisation": m.location,
   "FM_TageNachDiagnose": DaysSinceDiagnosis(m.date)
  }

define surgeryOutput:
 [Surgery] s
  return all {
   "TumorId": Tumor.tumorId,
   "OP_Intention": s.intention,
   "OP_TageNachDiagnose": DaysSinceDiagnosis(s.date)
  }

define radioOutput:
 Flatten(from
    [Radiotherapy] r
    where r.radiations is not null
    return all(from
       r.radiations rad
       return all {
        "TumorId": Tumor.tumorId,
        "ST_Intention": r.intention,
```

```
        "STBeginn_TageNachDiagnose": DaysSinceDiagnosis(rad.startDate)
      }
    )
 )

define systemicOutput:
 Flatten(from
    [SystemicTherapy] s
    where s.therapies is not null
    return all(from
      s.therapies t
      where t is not null
      return all {
        "TumorId": Tumor.tumorId,
        "SYT_Intention": s.intention,
        "SYTBeginn_TageNachDiagnose": DaysSinceDiagnosis(start of s.period),
        "SYT_ART": t
      }
    )
 )
```

4.3 SQL Abfrage KLast

Im Folgenden ist das im Rahmen der Evaluation für PMAW in der KLast verwendete Abfrageskript wiedergegeben, deren Ergebnis als Goldstandard gilt. Das Skript gibt die Filterkriterien wieder:

```
-- Tumor
select distinct tumor.TumorId,
   tumor.PatientId,
   tumor.DiagnosisAge as Diagnosealter,
   tumor.Gender as Geschlecht,
   substring (tumor.PrimaryTumorIcdCode, 1, 3) as Primaerdiagnose,
   YEAR(tumor.DiagnosisDate) as Diagnosejahr,
   tumor.MorphologyCode as Morphologie_Code,
   tumor.Side,
   tumor.Grading,
   tumor.OverallPerformanceStatus as Allgemeiner_Leistungszustand,
   [BestOfTnmT],
   [BestOfTnmN],
   [BestOfTnmM],
   DATEDIFF(dd, tumor.DiagnosisDate, D.DeathDate) as Vitalstatus_TageNachDiagnose,
   D.DeathTumorRelated as Tod_Tumorbedingt
FROM [fact_tumor] tumor
   left outer join [fact_tumor_staging] tumorStaging ON tumor.TumorId = tumorStaging.TumorId
   left outer join [fact_death] D on tumor.PatientId = D.PatientId
   left outer join [fact_vital_status] VS on tumor.PatientId = VS.PatientId
where tumor.TumorId in (
      Select DISTINCT tumor.TumorId
```

```
    FROM [fact_tumor] tumor
      full outer join [fact_tumor_staging] tumorStaging ON tumor.TumorId = tumorStaging.TumorId
      full outer join [fact_systemic_therapy] SYST on tumor.TumorId = SYST.TumorId
    WHERE (PrimaryTumorIcdCode LIKE 'C50%')
      AND BestOfTnmM LIKE '%M1%'
      AND (
        Gender like 'w'
        or Gender like 'm'
      )
      and MorphologyCode not like '%/6'
      and DiagnosisDate > '2018-07-01 00:00:00'
      and SYST.Intention like 'P'
  )
order by tumor.patientid,
  tumor.TumorId,
  Vitalstatus_TageNachDiagnose;
-- Fernmetastasen
select M.TumorId,
  [Localisation] as FM_Lokalisation,
  DATEDIFF(dd, tumor.DiagnosisDate, DateDistantMetastasis) as FM_TageNachDiagnose
FROM [fact_tumor] tumor,
  [fact_distant_metastasis] M
where tumor.TumorId = M.TumorId
  and M.tumorid in (
    Select DISTINCT tumor.TumorId
    FROM [fact_tumor] tumor
      full outer join [fact_tumor_staging] tumorStaging ON tumor.TumorId = tumorStaging.TumorId
      full outer join [fact_systemic_therapy] SYST on tumor.TumorId = SYST.TumorId
    WHERE (PrimaryTumorIcdCode LIKE 'C50%')
      AND BestOfTnmM LIKE '%M1%'
      AND (
        Gender like 'w'
        or Gender like 'm'
      )
      and MorphologyCode not like '%/6'
      and DiagnosisDate > '2018-07-01 00:00:00'
      and SYST.Intention like 'P'
  )
order by patientid,
  M.tumorid;
-- Operation
select OP.TumorId,
  [Intention] as OP_Intention,
  DATEDIFF(dd, tumor.DiagnosisDate, SurgeryDate) as OP_TageNachDiagnose
FROM [fact_tumor] tumor,
  [fact_surgery] OP
where tumor.TumorId = OP.TumorId
  and OP.tumorid in (
    Select DISTINCT tumor.TumorId
    FROM [fact_tumor] tumor
```

```
        full outer join [fact_tumor_staging] tumorStaging ON tumor.TumorId = tumorStaging.TumorId
        full outer join [fact_systemic_therapy] SYST on tumor.TumorId = SYST.TumorId
      WHERE (PrimaryTumorIcdCode LIKE 'C50%')
        AND BestOfTnmM LIKE '%M1%'
        AND (
          Gender like 'w'
          or Gender like 'm'
        )
        and MorphologyCode not like '%/6'
        and DiagnosisDate > '2018-07-01 00:00:00'
        and SYST.Intention like 'P'
    )
order by patientid,
    OP.tumorid;
-- Radiotherapie
select ST.TumorId,
    ST.Intention as ST_Intention,
    DATEDIFF(dd, tumor.DiagnosisDate, STGST.StartDate) as STBeginn_TageNachDiagnose
FROM [fact_tumor] tumor,
    [fact_radiotherapy] ST,
    [fact_radiotherapy_radiation] STGST
where tumor.tumorid = ST.TumorId
    and ST.RadiotherapyId = STGST.RadiotherapyId
    and ST.tumorid in (
      Select DISTINCT tumor.tumorid
      FROM [fact_tumor] tumor
        full outer join [fact_tumor_staging] tumorStaging ON tumor.tumorid = tumorStaging.TumorId
        full outer join [fact_systemic_therapy] SYST on tumor.tumorid = SYST.TumorId
      WHERE (PrimaryTumorIcdCode LIKE 'C50%')
        AND BestOfTnmM LIKE '%M1%'
        AND (
          Gender like 'w'
          or Gender like 'm'
        )
        and MorphologyCode not like '%/6'
        and DiagnosisDate > '2018-07-01 00:00:00'
        and SYST.Intention like 'P'
    )
order by ST.tumorid;
-- Systemische Therapie
select SYT.TumorId,
    [Intention] as SYT_Intention,
    DATEDIFF(dd, tumor.DiagnosisDate, BeginDate) as SYTBeginn_TageNachDiagnose,
    [TypeOfSystemicTherapy] as SYT_Art
FROM [fact_tumor] tumor,
    [fact_systemic_therapy] SYT,
    [fact_systemic_therapy_type] SYTA
where tumor.TumorId = SYT.TumorId
    and SYT.SystemicTherapyId = SYTA.SystemicTherapyId
    and SYT.tumorid in (
```

```
    Select DISTINCT tumor.TumorId
    FROM [fact_tumor] tumor
      full outer join [fact_tumor_staging] tumorStaging ON tumor.TumorId = tumorStaging.TumorId
      full outer join [fact_systemic_therapy] SYST on tumor.TumorId = SYST.TumorId
    WHERE (PrimaryTumorIcdCode LIKE 'C50%')
      AND BestOfTnmM LIKE '%M1%'
      AND (
        Gender like 'w'
        or Gender like 'm'
      )
      and MorphologyCode not like '%/6'
      and DiagnosisDate > '2018-07-01 00:00:00'
      and SYST.Intention like 'P'
  )
order by SYT.tumorid;
```

4.4 Ergebnisse

Es wurde die gleiche Anzahl an Fällen exportiert und es handelte sich um die gleichen Fälle. Es gab keine Abweichungen zwischen den Ergebnissen der PMAW CQL Anfragen und den SQL Anfragen der KLast.

Anfrage	**Fälle für den Datenexport**
CQL	1206
SQL KLast	1206

5 Routineauswertung 4a

In diesem Abschnitt wird exemplarisch aus dem Bereich „Routineauswertungen“ der Evaluation für das Fallbeispiel 4a die konkrete Umsetzung in eine CQL Abfrage mittels PMAW und ein Aufbereitungsskript in R abgedruckt.

Krebsregister veröffentlichen regelmäßig aggregierte Krebsdaten. Diese Routineauswertungen dienen der Berichtspflicht gegenüber der Öffentlichkeit, der Politik und Gesundheitsplanung. Im Bereich *Routineauswertung* wurden exemplarisch zwei Seiten des interaktiven Onlineberichts der KLast mit CQL abgefragt.

Der interaktive Online-Bericht der KLast bietet einen umfassenden Überblick über die Krebssituation in Niedersachsen. Für die Evaluation wurden zwei Seiten des Jahresberichts mit CQL analysiert: eine Seite zu neuen Krebsfällen (nach Diagnosegruppe, Geschlecht, Diagnosejahr und Wohnort – Fallbeispiel 4a Inzidenz) und eine Seite zu Operationen (mit zusätzlichen Merkmalen wie UICC-Stadium und Zeit bis zur ersten Operation – Fallbeispiel 4b Operation). Die Daten wurden in CQL gefiltert, als JSON exportiert und in R weiterverarbeitet, um die erforderlichen Kennzahlen zu berechnen. Für beide Anfragen bestand die untersuchte Kohorte aus 152.590 Krebsfällen. Für beide Anfragen stimmten die Ergebnisse mit den

bisherigen Methoden der KLast (Erstellung von MDX-Abfragen in einem Data Warehouse und anschließende Verarbeitung mit einem C#-Tool) exakt überein.

5.1 CQL Abfrage PMAW

Im Folgenden ist das im Rahmen der Evaluation für PMAW in CQL geschriebene Abfrageskript wiedergegeben:

```
using SepamimModel

context Tumor

define function StartsWithAny(str String, starts List<String>):
 AnyTrue(starts s
    return StartsWith(str, s)
 )

define function IcdGroup(icdCode String):
 case
  when StartsWithAny(icdCode, { 'C02', 'C03', 'C04', 'C05', 'C06', 'C00.3', 'C00.4', 'C00.5', 'C00.6', 'C00.7',
'C00.8', 'C00.9' })
    or icdCode = 'C00' then 'C00, C02-C06 ohne C00.0-C00.2'
  when StartsWithAny(icdCode, { 'C07', 'C08' }) then 'C07, C08'
  when StartsWithAny(icdCode, { 'C01', 'C09', 'C10', 'C11', 'C12', 'C13', 'C14.0', 'C14.2' })
    or icdCode = 'C14' then 'C01, C09-C14 ohne C14.8'
  when StartsWith(icdCode, 'C15') then 'C15'
  when StartsWith(icdCode, 'C16') then 'C16'
  when StartsWith(icdCode, 'C17') then 'C17'
  when StartsWith(icdCode, 'C18') then 'C18'
  when StartsWithAny(icdCode, { 'C19', 'C20' }) then 'C19, C20'
  when StartsWith(icdCode, 'C21') then 'C21'
  when StartsWith(icdCode, 'C22') then 'C22'
  when StartsWithAny(icdCode, { 'C23', 'C24' }) then 'C23-C24'
  when StartsWith(icdCode, 'C25') then 'C25'
  when StartsWith(icdCode, 'C26') then 'C26'
  when StartsWithAny(icdCode, { 'C30', 'C31' }) then 'C30, C31'
  when StartsWith(icdCode, 'C32') then 'C32'
  when StartsWith(icdCode, 'C33') then 'C33'
  when StartsWith(icdCode, 'C34') then 'C34'
  when StartsWithAny(icdCode, { 'C37', 'C38', 'C39' }) then 'C37-C39'
  when StartsWithAny(icdCode, { 'C40', 'C41' }) then 'C40, C41'
  when StartsWith(icdCode, 'C43') then 'C43'
  when StartsWith(icdCode, 'C45') then 'C45'
  when StartsWithAny(icdCode, { 'C46', 'C47', 'C48', 'C49' }) then 'C46-C49'
  when StartsWith(icdCode, 'C50') then 'C50'
  when StartsWith(icdCode, 'C51') then 'C51'
  when StartsWithAny(icdCode, { 'C52', 'C57', 'C58' }) then 'C52, C57-C58'
  when StartsWith(icdCode, 'C53') then 'C53'
  when StartsWithAny(icdCode, { 'C54', 'C55' }) then 'C54, C55'
```

```
    when StartsWith(icdCode, 'C56') then 'C56'
    when StartsWith(icdCode, 'C60') then 'C60'
    when StartsWith(icdCode, 'C61') then 'C61'
    when StartsWith(icdCode, 'C62') then 'C62'
    when StartsWith(icdCode, 'C64') then 'C64'
    when StartsWithAny(icdCode, { 'C65', 'C66', 'C68' }) then 'C65, C66, C68'
    when StartsWithAny(icdCode, { 'C67', 'D09.0', 'D41.4' }) then 'C67, D09.0, D41.4'
    when StartsWith(icdCode, 'C69') then 'C69'
    when StartsWith(icdCode, 'C70') then 'C70'
    when StartsWith(icdCode, 'C71') then 'C71'
    when StartsWith(icdCode, 'C72') then 'C72'
    when StartsWith(icdCode, 'C73') then 'C73'
    when StartsWithAny(icdCode, { 'C74', 'C75' }) then 'C74, C75'
    when StartsWithAny(icdCode, { 'C76', 'C80' }) then 'C76, C80'
    when StartsWith(icdCode, 'C81') then 'C81'
    when StartsWith(icdCode, 'C82') then 'C82'
    when StartsWith(icdCode, 'C83.3') then 'C83.3'
    when StartsWith(icdCode, 'C84') then 'C84'
    when StartsWith(icdCode, 'C85') then 'C85'
    when StartsWith(icdCode, 'C86') then 'C86'
    when StartsWith(icdCode, 'C88') then 'C88'
    when StartsWith(icdCode, 'C90') then 'C90'
    when StartsWith(icdCode, 'C91.1') then 'C91.1'
    when StartsWithAny(icdCode, { 'C91.0', 'C91.8' }) then 'C91.0, C91.8'
    when StartsWith(icdCode, 'C92.1') then 'C92.1'
    when StartsWithAny(icdCode, { 'C92.0', 'C92.3', 'C92.4', 'C92.5', 'C92.6', 'C92.8', 'C93.0', 'C94.0', 'C94.2',
'C94.3', 'C94.4' }) then 'C92.0, C92.3-C92.6, C92.8, C93.0, C94.0, C94.2-C94.4'
    when StartsWithAny(icdCode, { 'C92.2', 'C93.1', 'C93.3' }) then 'C92.2, C93.1, C93.3'
    when StartsWith(icdCode, 'D03') then 'D03'
    when StartsWith(icdCode, 'D05') then 'D05'
    when StartsWith(icdCode, 'D06') then 'D06'
    when StartsWithAny(icdCode, { 'D32', 'D33' }) then 'D32, D33'
    else 'other' end

define "Initiale Population":
  Tumor.numberOfClinicalReports > 0
    and not Tumor.rued
    and ( EndsWith(Tumor.morphologyCodeIcdO, '/0')
      or EndsWith(Tumor.morphologyCodeIcdO, '/1')
      or EndsWith(Tumor.morphologyCodeIcdO, '/2')
      or EndsWith(Tumor.morphologyCodeIcdO, '/3')
      or EndsWith(Tumor.morphologyCodeIcdO, '/9')
      or Tumor.morphologyCodeIcdO = 'X'
      or PositionOf('/', Tumor.morphologyCodeIcdO) = - 1
    )
    and Tumor.icdCode in { 'C00', 'C00.0', 'C00.1', 'C00.2', 'C00.3', 'C00.4', 'C00.5', 'C00.6', 'C00.8', 'C00.9',
'C00-C14', 'C01', 'C02', 'C02.0', 'C02.1', 'C02.2', 'C02.3', 'C02.4', 'C02.8', 'C02.9', 'C03', 'C03.0', 'C03.1',
'C03.9', 'C04', 'C04.0', 'C04.1', 'C04.8', 'C04.9', 'C05', 'C05.0', 'C05.1', 'C05.2', 'C05.8', 'C05.9', 'C06', 'C06.0',
'C06.1', 'C06.2', 'C06.8', 'C06.9', 'C07', 'C08', 'C08.0', 'C08.1', 'C08.8', 'C08.9', 'C09', 'C09.0', 'C09.1', 'C09.8',
'C09.9', 'C10', 'C10.0', 'C10.1', 'C10.2', 'C10.3', 'C10.4', 'C10.8', 'C10.9', 'C11', 'C11.0', 'C11.1', 'C11.2',
```

'C11.3', 'C11.8', 'C11.9', 'C12', 'C13', 'C13.0', 'C13.1', 'C13.2', 'C13.8', 'C13.9', 'C14', 'C14.0', 'C14.2', 'C14.8', 'C15', 'C15.0', 'C15.1', 'C15.2', 'C15.3', 'C15.4', 'C15.5', 'C15.8', 'C15.9', 'C15-C26', 'C16', 'C16.0', 'C16.1', 'C16.2', 'C16.3', 'C16.4', 'C16.5', 'C16.6', 'C16.8', 'C16.9', 'C17', 'C17.0', 'C17.1', 'C17.2', 'C17.3', 'C17.8', 'C17.9', 'C18', 'C18.0', 'C18.1', 'C18.2', 'C18.3', 'C18.4', 'C18.5', 'C18.6', 'C18.7', 'C18.8', 'C18.9', 'C19', 'C20', 'C21', 'C21.0', 'C21.1', 'C21.2', 'C21.8', 'C22', 'C22.0', 'C22.1', 'C22.2', 'C22.3', 'C22.4', 'C22.7', 'C22.9', 'C23', 'C24', 'C24.0', 'C24.1', 'C24.8', 'C24.9', 'C25', 'C25.0', 'C25.1', 'C25.2', 'C25.3', 'C25.4', 'C25.7', 'C25.8', 'C25.9', 'C26', 'C26.0', 'C26.1', 'C26.8', 'C26.9', 'C30', 'C30.0', 'C30.1', 'C30-C39', 'C31', 'C31.0', 'C31.1', 'C31.2', 'C31.3', 'C31.8', 'C31.9', 'C32', 'C32.0', 'C32.1', 'C32.2', 'C32.3', 'C32.8', 'C32.9', 'C33', 'C34', 'C34.0', 'C34.1', 'C34.2', 'C34.3', 'C34.8', 'C34.9', 'C37', 'C38', 'C38.0', 'C38.1', 'C38.2', 'C38.3', 'C38.4', 'C38.8', 'C39', 'C39.0', 'C39.8', 'C39.9', 'C40', 'C40.0', 'C40.1', 'C40.2', 'C40.3', 'C40.8', 'C40.9', 'C40-C41', 'C41', 'C41.0', 'C41.01', 'C41.02', 'C41.1', 'C41.2', 'C41.3', 'C41.30', 'C41.31', 'C41.32', 'C41.4', 'C41.8', 'C41.9', 'C43', 'C43.0', 'C43.1', 'C43.2', 'C43.3', 'C43.4', 'C43.5', 'C43.6', 'C43.7', 'C43.8', 'C43.9', 'C45', 'C45.0', 'C45.1', 'C45.2', 'C45.7', 'C45.9', 'C45-C49', 'C46', 'C46.0', 'C46.1', 'C46.2', 'C46.3', 'C46.7', 'C46.8', 'C46.9', 'C47', 'C47.0', 'C47.1', 'C47.2', 'C47.3', 'C47.4', 'C47.5', 'C47.6', 'C47.8', 'C47.9', 'C48', 'C48.0', 'C48.1', 'C48.2', 'C48.8', 'C49', 'C49.0', 'C49.1', 'C49.2', 'C49.3', 'C49.4', 'C49.5', 'C49.6', 'C49.8', 'C49.9', 'C50', 'C50.0', 'C50.1', 'C50.2', 'C50.3', 'C50.4', 'C50.5', 'C50.6', 'C50.8', 'C50.9', 'C50-C50', 'C51', 'C51.0', 'C51.1', 'C51.2', 'C51.8', 'C51.9', 'C51-C58', 'C52', 'C53', 'C53.0', 'C53.1', 'C53.8', 'C53.9', 'C54', 'C54.0', 'C54.1', 'C54.2', 'C54.3', 'C54.8', 'C54.9', 'C55', 'C56', 'C57', 'C57.0', 'C57.1', 'C57.2', 'C57.3', 'C57.4', 'C57.7', 'C57.8', 'C57.9', 'C58', 'C60', 'C60.0', 'C60.1', 'C60.2', 'C60.8', 'C60.9', 'C60-C63', 'C61', 'C62', 'C62.0', 'C62.1', 'C62.9', 'C63', 'C63.0', 'C63.1', 'C63.2', 'C63.7', 'C63.8', 'C63.9', 'C64', 'C64-C68', 'C65', 'C66', 'C67', 'C67.0', 'C67.1', 'C67.2', 'C67.3', 'C67.4', 'C67.5', 'C67.6', 'C67.7', 'C67.8', 'C67.9', 'C68', 'C68.0', 'C68.1', 'C68.8', 'C68.9', 'C69', 'C69.0', 'C69.1', 'C69.2', 'C69.3', 'C69.4', 'C69.5', 'C69.6', 'C69.8', 'C69.9', 'C69-C72', 'C70', 'C70.0', 'C70.1', 'C70.9', 'C71', 'C71.0', 'C71.1', 'C71.2', 'C71.3', 'C71.4', 'C71.5', 'C71.6', 'C71.7', 'C71.8', 'C71.9', 'C72', 'C72.0', 'C72.1', 'C72.2', 'C72.3', 'C72.4', 'C72.5', 'C72.8', 'C72.9', 'C73', 'C73-C75', 'C74', 'C74.0', 'C74.1', 'C74.9', 'C75', 'C75.0', 'C75.1', 'C75.2', 'C75.3', 'C75.4', 'C75.5', 'C75.8', 'C75.9', 'C76', 'C76.0', 'C76.1', 'C76.2', 'C76.3', 'C76.4', 'C76.5', 'C76.7', 'C76.8', 'C80', 'C80.0', 'C80.9', 'C81', 'C81.0', 'C81.1', 'C81.2', 'C81.3', 'C81.4', 'C81.7', 'C81.9', 'C81-C96', 'C82', 'C82.0', 'C82.1', 'C82.2', 'C82.3', 'C82.4', 'C82.5', 'C82.6', 'C82.7', 'C82.9', 'C83', 'C83.0', 'C83.1', 'C83.3', 'C83.5', 'C83.7', 'C83.8', 'C83.9', 'C84', 'C84.0', 'C84.1', 'C84.4', 'C84.5', 'C84.6', 'C84.7', 'C84.8', 'C84.9', 'C85', 'C85.1', 'C85.2', 'C85.7', 'C85.9', 'C86', 'C86.0', 'C86.1', 'C86.2', 'C86.3', 'C86.4', 'C86.5', 'C86.6', 'C88', 'C88.0', 'C88.00', 'C88.01', 'C88.2', 'C88.20', 'C88.21', 'C88.3', 'C88.30', 'C88.31', 'C88.4', 'C88.40', 'C88.41', 'C88.7', 'C88.70', 'C88.71', 'C88.9', 'C88.90', 'C88.91', 'C90', 'C90.0', 'C90.00', 'C90.01', 'C90.1', 'C90.10', 'C90.11', 'C90.2', 'C90.20', 'C90.21', 'C90.3', 'C90.30', 'C90.31', 'C91', 'C91.0', 'C91.00', 'C91.01', 'C91.1', 'C91.10', 'C91.11', 'C91.3', 'C91.30', 'C91.31', 'C91.4', 'C91.40', 'C91.41', 'C91.5', 'C91.50', 'C91.51', 'C91.6', 'C91.60', 'C91.61', 'C91.7', 'C91.70', 'C91.71', 'C91.8', 'C91.80', 'C91.81', 'C91.9', 'C91.90', 'C91.91', 'C92', 'C92.0', 'C92.00', 'C92.01', 'C92.1', 'C92.10', 'C92.11', 'C92.2', 'C92.20', 'C92.21', 'C92.3', 'C92.30', 'C92.31', 'C92.4', 'C92.40', 'C92.41', 'C92.5', 'C92.50', 'C92.51', 'C92.6', 'C92.60', 'C92.61', 'C92.7', 'C92.70', 'C92.71', 'C92.8', 'C92.80', 'C92.81', 'C92.9', 'C92.90', 'C92.91', 'C93', 'C93.0', 'C93.00', 'C93.01', 'C93.1', 'C93.10', 'C93.11', 'C93.3', 'C93.30', 'C93.31', 'C93.7', 'C93.70', 'C93.71', 'C93.9', 'C93.90', 'C93.91', 'C94', 'C94.0', 'C94.00', 'C94.01', 'C94.2', 'C94.20', 'C94.21', 'C94.3', 'C94.30', 'C94.31', 'C94.4', 'C94.40', 'C94.41', 'C94.6', 'C94.60', 'C94.61', 'C94.7', 'C94.70', 'C94.71', 'C94.8!', 'C95', 'C95.0', 'C95.00', 'C95.01', 'C95.1', 'C95.10', 'C95.11', 'C95.7', 'C95.70', 'C95.71', 'C95.8!', 'C95.9', 'C95.90', 'C95.91', 'C96', 'C96.0', 'C96.2', 'C96.4', 'C96.5', 'C96.6', 'C96.7', 'C96.8', 'C96.9', 'D00', 'D00.0', 'D00.1', 'D00.2', 'D00-D09', 'D01', 'D01.0', 'D01.1', 'D01.2', 'D01.3', 'D01.4', 'D01.5', 'D01.7', 'D01.9', 'D02', 'D02.0', 'D02.1', 'D02.2', 'D02.3', 'D02.4', 'D03', 'D03.0', 'D03.1', 'D03.2', 'D03.3', 'D03.4', 'D03.5', 'D03.6', 'D03.7', 'D03.8', 'D03.9', 'D05', 'D05.0', 'D05.1', 'D05.7', 'D05.9', 'D06', 'D06.0', 'D06.1', 'D06.7', 'D06.9', 'D07', 'D07.0', 'D07.1', 'D07.2', 'D07.3', 'D07.4', 'D07.5', 'D07.6', 'D09', 'D09.0', 'D09.1', 'D09.2', 'D09.3', 'D09.7', 'D09.9', 'D32', 'D32.0', 'D32.1', 'D32.9', 'D33', 'D33.0', 'D33.1', 'D33.2', 'D33.3', 'D33.4', 'D33.7', 'D33.9', 'D35.2', 'D35.3', 'D35.4', 'D39.1', 'D41.4', 'D42', 'D42.0', 'D42.1', 'D42.9', 'D43', 'D43.0', 'D43.1', 'D43.2', 'D43.3', 'D43.4', 'D43.7', 'D43.9', 'D44.3', 'D44.4', 'D44.5', 'D45', 'D46', 'D46.0', 'D46.1', 'D46.2', 'D46.4', 'D46.5', 'D46.6', 'D46.7', 'D46.9', 'D47.1', 'D47.3', 'D47.4', 'D47.5' }

```
define "Weitere Filter":
 "Initiale Population"
  and year from Tumor.dateOfDiagnosis in { 2019, 2020, 2021 }

define Output:
 {
  IcdGroup: IcdGroup(Tumor.icdCode),
  Year: year from Tumor.dateOfDiagnosis,
  Sex: Tumor.sex,
  Population: 'ALL'
 }
```

5.2 Aufbereitungs-Skript in R

Im Folgenden ist das im Rahmen der Evaluation für PMAW in R geschriebene Skript wiedergegeben, welches das JSON Ergebnis der CQL Anfrage in ein CSV Format zum Abgleich mit dem KLast Ergebnis konvertiert:

```
library(dplyr)
library(tidyr)
library(jsonlite)
library(data.table)
library(readr)
cql_df <- fromJSON("./general_overview.json") |>
  mutate(across(where(is.numeric), as.character))

dt <- cql_df |>
  data.table() |>
  groupingsets(
    j = .(c = .N),
    by = c("Year", "IcdGroup", "Sex", "Population"),
    sets = list(
      c("Year", "IcdGroup", "Sex", "Population"),
      c("Year", "IcdGroup", "Population"),
      c("Year", "Sex", "Population"),
      c("Year", "Population")
    )
  ) |>
  replace_na(list(IcdGroup = "all", Sex = "ALL", Fallzahl.X=0)) |>
  filter(Sex %in% c("M", "W", "ALL")) |>
  pivot_wider(names_from = Sex, values_from = c) |>
  relocate(
    DY = Year,
    DG = IcdGroup,
    GG = Population,
    `Fallzahl M` = M,
    `Fallzahl W` = W,
```

```
    Gesamt = ALL
  ) |>
  replace_na(list(`Fallzahl M`=0, `Fallzahl W` = 0, Gesamt=0)) |>

  arrange(pick(where(is.character)))

result <- read_csv("AI_HaeufigsteKE.csv", col_types = "ccciii") |>
  filter(GG == "ALL") |>
  arrange(pick(where(is.character))) |>
  as_tibble()


all.equal(result, dt)
all_equal(result, dt)

#print(result[59,], width=1000)
#print(dt[59,], width=1000)
#print(result[60,], width=1000)
#print(dt[60,], width=1000)
```

5.3 Ergebnisse

Für beide Anfragen der Routineauswertung der KLast bestand die untersuchte Kohorte aus 152.590 Krebsfällen. Für beide Anfragen stimmten die Ergebnisse mit den bisherigen Methoden der KLast (Erstellung von MDX-Abfragen in einem Data Warehouse und anschließende Verarbeitung mit einem C#-Tool) exakt überein.
Das Tooling der KLast für den Goldstandard ist proprietär und deshalb nicht in der Softwaresammlung enthalten. Die Evaluationsskripte für PMAW sind enthalten.

11. MÄRZ 2025

Version v1.1

Anforderungsanalyse Pattern Matching Abfragewerkzeug (MS 4)

KOLJA BLOHM, DAVID KORFKAMP
OFFIS E.V. – INSTITUT FÜR INFORMATIK
Escherweg 2, 26121 Oldenburg

Inhaltsverzeichnis

1. Einleitung

Das Pattern Matching Abfragewerkzeug (PMAW) soll die Mitarbeiter*innen des Krebsregisters NRW bei der Operationalisierung von Daten zu Behandlungs- und Erkrankungsverläufen aus dem oBDS-Basisdatensatz unterstützen. Hierzu sollen Kohorten anhand der Patient*innen- und Diagnosedaten sowie der Verlaufsdaten von Patient*innen gebildet werden. Die Verlaufsdaten sollen anhand eines vorgegebenen Suchmusters gefiltert werden. Ein Suchmuster für den Verlauf besteht dabei aus mehreren Ereignissen die während der Behandlung aufgetreten sind, sowie der zeitlichen Beziehung zwischen den Ereignissen. Die Suche anhand von bestimmten Krankheitsverläufen kann dabei mehreren Zwecken dienen, wie z.B. der Überprüfung der Plausibilität der Daten, der Überprüfung von Leitlinien bzw. Behandlungsplänen oder der Identifizierung von Patient*innenkohorten für die weitere Analyse.

Der Anforderungskatalog mit Stand 30.09.2021 wurde nach Projektende um eine Bewertung des Erfüllungsgrads erweitert. Diese wurde zu den einzelnen Anforderungen notiert. Alle Muss-Anforderungen wurden vollständig umgesetzt, einzelne Kann- und Soll-Anforderungen wurden nur teilweise bzw. wenige gar nicht umgesetzt.

2. Einsatzumgebung

Das PMAW soll im klinischen Krebsregister Nordrhein-Westfalen zum Einsatz kommen. Das PMAW soll ausschließlich intern von Mitarbeiter*innen des Krebsregisters genutzt werden und richtet sich somit an Fachanwender*innen.

3. Datengrundlage

In diesem Abschnitt werden die für das PMAW benötigten Daten beschrieben. Die Grundlage bilden Best-Of Daten, die aus dem oBDS-Basisdatensatz abgeleitet wurden. Best-Of Daten sind für die Analyse bereinigte Datensätze, in denen Probleme in der Datenqualität, die beispielsweise dadurch entstehen können, dass widersprüchliche (z.B. Ende einer Therapie vor Beginn) oder unvollständige Meldungen (z.B. Operationsmeldungen ohne OPS-Codes) zu einem Tumor auftreten, behoben wurden. Der Best-Of Datensatz besteht aus Verlaufsdaten sowie Patienten- und Diagnosedaten.

Verlaufsdaten stellen Ereignisse dar, die im Krankheits- und Behandlungsverlauf eines Patienten auftreten. Hierunter fallen unter Anderem Operationen und andere Therapien aber auch Meldungen zum Status des Tumors, wie Metastasen oder Veränderungen der Tumorgröße. Bei den Ereignissen lässt sich zwischen Zeitpunkt-Ereignissen und Intervall-Ereignissen unterscheiden. Zeitpunkt-Ereignisse haben einen einzelnen Zeitstempel, der den Zeitpunkt beschreibt, an dem ein Ereignis aufgetreten ist. Die Dauer des Ereignisses ist hier nicht relevant. Ein Beispiel hierfür ist eine Operation des Tumors. Intervall-Ereignisse haben einen Start- und (optionalen) Endzeitpunkt, da hier die Dauer des Ereignisses relevant für die Analyse ist. Ein Beispiel hierfür ist eine Chemotherapie, die über mehrere Wochen ausgeführt wird.

Patienten- und Diagnosedaten sind verlaufsunabhängige Daten, die sich auf einen Patienten bzw. eine Tumorerkrankung als Ganzes beziehen. Beispiele hierfür sind das Geschlecht oder das Alter des Patienten sowie die Primärdiagnose. Diese Daten dienen vor allem dazu, bei der Suche anhand von Behandlungsverläufen die relevanten Tumore vorzufiltern. Die konkreten

Entitäten und Ereignistypen und deren Ausprägungen werden im Konzept zum Analysemodell definiert.

4. Anforderungen und ihr Erfüllungsgrad nach Projektende

Dieser Abschnitt enthält Anforderungen an das PMAW. Die Anforderungen sind in Kooperation zwischen dem LKR.NRW und dem OFFIS entstanden und bilden den derzeitigen Kenntnisstand ab.

Das PMAW soll es ermöglichen, Patienten anhand von Patienten- und Diagnosedaten sowie Verlaufsdaten zu filtern. Mit dem Filtern anhand von Patienten- und Diagnosedaten ist beispielsweise die Einschränkung auf Patienten mit einem bestimmten Geschlecht oder einer bestimmten Diagnose gemeint. Die Verlaufsdaten sollen anhand eines vorgegebenen Suchmusters gefiltert werden. Das Suchmuster beschreibt relevante Ereignisse, die aufgetreten sein müssen (z.B. eine Operation mit einem bestimmten OPS-Code), oder auch die zeitliche Beziehung zwischen Ereignissen (z.B. Operation fand 3 Wochen nach Diagnose statt). Jedes Ereignis besitzt einen Typ (z.B. Operation, Strahlentherapie). Der Ereignistyp definiert die Attribute, die ein Ereignis dieses Typs besitzt (z.B. OPS-Codes bei Operations-Ereignissen). Die konkreten Entitäten und Ereignistypen und deren Ausprägungen sind im Konzept zum Analysemodell beschrieben.

Die folgende Formulierung der Anforderungen orientiert sich an gängigen Anforderungsschablonen, wie sie z.B. in [Rupp], [Partsch] gefunden werden können. Die Wichtigkeit einer Anforderung wird über ein Schlüsselwort in der Formulierung bestimmt:

- *MUSS*-Anforderungen sind Anforderungen die verpflichtend umgesetzt werden müssen.
- *SOLL*-Anforderungen sind Anforderungen die nicht verpflichtend umgesetzt werden müssen, aber eine hohe Priorität aufweisen.
- *KANN*-Anforderungen sind Anforderungen die nicht verpflichtend umgesetzt werden müssen, und eine niedrige Priorität aufweisen.

Nach Projektende wurde eine Bewertung des Erfüllungsgrads unter den jeweiligen Anforderungen vorgenommen, eingeleitet mit dem Tag „[Erfüllungsgrad Projektende 2024]“. Die Aussage bezieht sich dabei immer auf den darüberstehenden Eintrag und i.d.R. alle Sub-Elemente. Auf Ausnahmen wird hingewiesen.

Funktionale Anforderungen

Anforderungen an die Filterung anhand der Patienten- und Diagnosedaten

- [Erfüllungsgrad Projektende 2024] Alle genannten Anforderungen wurden in PMAW umgesetzt und sind erfüllt.
- Patienten- und Diagnosedaten müssen sich anhand von Attributen, wie z.B. dem ICD-10 Code der Diagnose oder dem Geschlecht, filtern lassen
- In einer Anfrage müssen sich Patienten- und Diagnosedaten anhand mehrerer Attribute gleichzeitig filtern lassen. Diese müssen sich über boolesche Logik miteinander verknüpfen lassen

Anforderungen an die Definition von relevanten Ereignissen

- [Erfüllungsgrad Projektende 2024] Alle genannten Anforderungen wurden in PMAW umgesetzt und sind erfüllt.
- Ereignisse müssen sich anhand von Ereignistypen, wie z.B. Operation oder Strahlentherapie, filtern lassen
- Ereignisse müssen sich anhand von Attributen, wie z.B. dem OPS Code einer Operation, filtern lassen
- In einer Anfrage müssen sich Ereignisse anhand mehrerer Attribute gleichzeitig filtern lassen. Diese müssen sich über boolesche Logik miteinander verknüpfen lassen

Anforderungen an die Definition von zeitlichen Beziehungen zwischen Ereignissen

- [Erfüllungsgrad Projektende 2024] Alle genannten Anforderungen wurden in PMAW umgesetzt und sind erfüllt.
- Die zeitliche Beziehung zwischen zwei Ereignissen muss sich abbilden lassen
    - Alle Operatoren der Allen-Algebra ([Allen]) sollen sich abbilden lassen
    - Es soll möglich sein, die zeitliche Dauer zwischen zwei Ereignissen zu spezifizieren (z.B. Therapie A findet mindestens 3 Wochen vor Therapie B statt)
    - Es soll möglich sein, mehrere per ODER verknüpfte zeitliche Operatoren zwischen zwei Ereignissen zu definieren (z.B. Therapie A vor ODER während Therapie B)
- Die Konjunktion von Ereignispaaren soll spezifizierbar sein (z.B. Therapie A vor Therapie B UND Therapie A vor Therapie C)
- Die Disjunktion von Ereignispaaren soll spezifizierbar sein (z.B. Therapie A vor Therapie B ODER Therapie A vor Therapie C)
- Es soll möglich sein, das Nicht-Auftreten eines Ereignisses zu spezifizieren (Negation)
- Es kann möglich sein, das erste Auftreten eines Ereignisses zu spezifizieren
- Es kann möglich sein, das letzte Auftreten eines Ereignisses zu spezifizieren

Anforderungen an die Ergebnisse der Suche

[Erfüllungsgrad Projektende 2024] Alle genannten Muss-Anforderungen wurden in PMAW umgesetzt und sind erfüllt. Die Soll- und Kann-Anforderungen wurden weitestgehend umgesetzt. Bei nicht vollständig umgesetzten Anforderungen werden die Einschränkungen durch einen Hinweis erläutert.

- Es muss möglich sein, die Tumor IDs der gefundenen Tumore auszugeben
- Es soll möglich sein, die Patienten IDs der gefundenen Tumore auszugeben
- Es soll möglich sein, die Verläufe der gefundenen Tumore auszugeben
    - [Erfüllungsgrad Projektende 2024] Der Gesamtverlauf einzelner Tumore wird visualisiert. Für einen Teilverlauf muss der Endanwender diesen durch Filterung spezifizieren und kann das Ergebnis dann als JSON Objekt (einem Datenstandard für komplexere Datenstrukturen) exportieren und bei Bedarf weiterverarbeiten und visualisieren. Es ist also möglich, erfordert aber zusätzlichen Aufwand und CQL-Expertise und hat entsprechend eine „Bedienungshürde“ bzw. ist eher technisch versierteren Nutzern möglich.
    - Es soll möglich sein, den gesamten Verlauf auszugeben
    - Es soll möglich sein, einen Teilverlauf auszugeben

- Es kann eine Überlebenszeitanalyse auf den gefundenen Tumoren ausgeführt und ausgegeben werden
- Es kann möglich sein, häufige Folgeereignisse zu visualisieren
    - [Erfüllungsgrad Projektende 2024] Eine Auswertung häufiger Folgeereignisse wurde nicht umgesetzt. Durch die verwendeten Sankey Charts lässt sich aber die grundsätzliche Verteilung visualisieren.
- Es kann möglich sein, Aggregate über die gefundenen Daten zu berechnen und auszugeben
    - [Erfüllungsgrad Projektende 2024] Dies wurde in PMAW nicht umgesetzt. Die gefundenen Daten müssen exportiert und dann in einem separaten Tool weiterausgewertet werden.
- Es soll möglich sein die Ergebnisse zu exportieren
    - Es soll möglich sein, die Tumor- und Patienten-IDs der gefundenen Ergebnisse zu exportieren
    - Es kann möglich sein, die Metadaten der Anfrage zu exportieren
        - [Erfüllungsgrad Projektende 2024] Ein vollständiger Export der Metadaten der Ergebnisse ist nicht möglich
    - Der Export soll im CSV-Format erfolgen
        - [Erfüllungsgrad Projektende 2024] Es ist teilweise umgesetzt: Ein Export im CSV Format ist für die Patienten-IDs, Tumor-IDs und weitere Attribute der Diagnose möglich. Ein Export ganzer Behandlungsverläufe, die aus verschachtelten und komplexen Ereignissen bestehen, istnur als JSON Objekt (einem Datenstandard für komplexere Datenstrukturen) möglich.

Anforderungen an die Benutzeroberfläche

[Erfüllungsgrad Projektende 2024] Alle genannten Soll-Anforderungen wurden in PMAW umgesetzt und sind erfüllt. Die Kann-Anforderungen wurden nicht alle umgesetzt. Bei nicht vollständig umgesetzten Anforderungen werden die Einschränkungen durch einen Hinweis erläutert.

- Das PMAW soll über eine grafische Benutzeroberfläche bedienbar sein
- Die Benutzeroberfläche soll es ermöglichen, Anfragen textuell zu definieren
- Die Benutzeroberfläche kann es ermöglichen, Anfragen grafisch zu definieren
    - [Erfüllungsgrad Projektende 2024] Anfragen können nur textuell und nicht grafisch definiert werden.
- Die Benutzeroberfläche kann Vorschläge für verfügbare Attribute oder valide Filterwerte machen
    - [Erfüllungsgrad Projektende 2024] Ein Vorschlagssystem wurde nicht umgesetzt. Jedoch wird das Schema des Datenmodells visualisiert aus dem die verfügbaren Attribute ersichtlich sind.

Nicht-funktionale Anforderungen

Bedienbarkeit durch Anwender*innen mit wenig technischem Hintergrund

- [Erfüllungsgrad Projektende 2024] Der Erfüllungsgrad für die nicht-funktionalen Anforderungen lässt sich nur näherungsweise angeben, da keine Nutzer*innen-Studie im Rahmen der Evaluation stattfand.
- Das PMAW soll von Anwender*innen mit wenig technischem Hintergrund selbstständig bedienbar sein
    - [Erfüllungsgrad Projektende 2024] Alle Nutzer*innen in den an der Evaluation beteiligten Krebsregistern konnten die Anwendung bedienen.
- Vorhandene Anfragen sollen für Anwender*innen mit wenig technischem Hintergrund verständlich sein
    - [Erfüllungsgrad Projektende 2024] Die verwendete Sprache zur Definition von Abfragen ist CQL. CQL ist explizit eine vereinfachende Anfragesprache für Domänenexperten im Gesundheitswesen mit wenig IT- Verständnis. Hierfür wurde CQL entwickelt. Im Rahmen der Evaluation wurde CQL mit den Expert*innen diskutiert. Das Verständnis der Anfragen ist bei CQL grundsätzlich gegeben und in den Diskussionen wurden die CQL-Anfragen als leichter als SQL eingestuft.
- Anfragen sollen von Anwender*innen mit wenig technischem Hintergrund selbstständig formuliert werden können
    - [Erfüllungsgrad Projektende 2024] Dieser Punkt verbleibt unklar, da es keine Anwender*innen außerhalb des Projekts gab, die eigene Anfragen neu definieren mussten. Wie beim vorherigen Punkt dürfte CQL leichter zu definieren sein als SQL, es bedarf aber trotzdem einer Schulung und für eine verlässliche Aussage einer tiefergehenden Evaluation.

5. Referenzen

[Rupp] Rupp, C. und Joppich, R. (2014). „Anforderungsschablonen – der MASTeR-Plan für gute Anforderungen“. In: Requirements-Engineering und -Management. Aus der Praxis von klassisch bis agil. Herausgeber: Rupp, C. und die SOPHISTen. Hanser, S. 215–246.

[Partsch] Partsch, H. (2010). „Requirements-Engineering systematisch. Modellbildung für softwaregestützte Systeme“. 2. Auflage. Springer-Verlag Berlin Heidelberg.

[Allen] Allen, James F. (1983). „Maintaining knowledge about temporal intervals.“ Communications of the ACM 26.11 (1983): S. 832-843.

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

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

Inhalt

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.

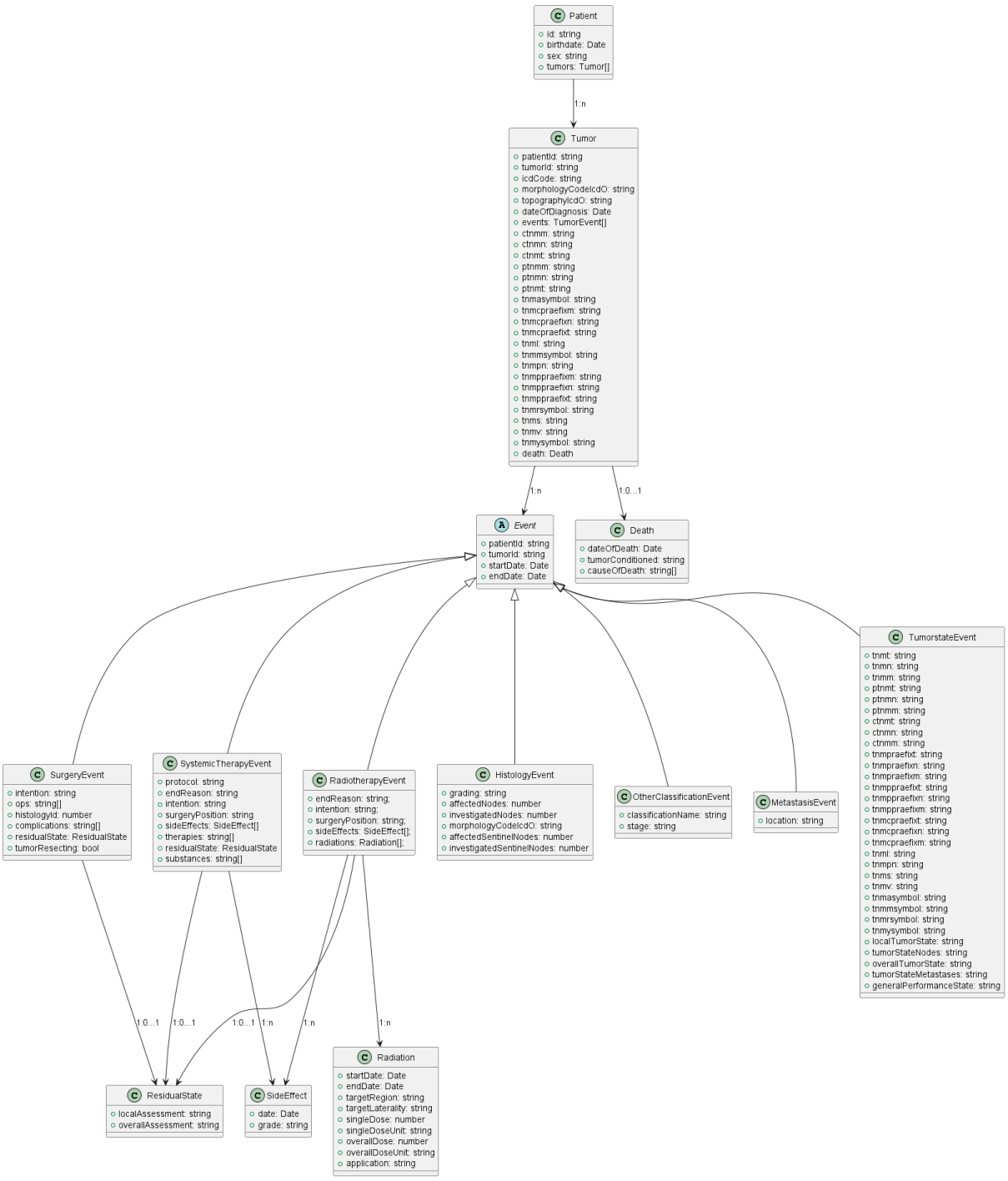


Abbildung 1 Entitäten des Analysemodells

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[1] 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.

[1] https://github.com/cqframework/clinical_quality_language

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

Gefördert vom G-BA Innovationsfond
FKZ: 01VSF20018

KONZEPTION EINES DATA MINING WERKZEUGS

13. DEZEMBER 2022

Version v1.0

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

GESAMTPROJEKTLEITUNG	PROF. DR.-ING. ANDREAS HEIN OFFIS – INSTITUT FÜR INFORMATIK
PROJEKTLEITUNG DER EVALUATION	FLORIAN OESTERLING LANDESKREBSREGISTER NRW, BOCHUM
KONSORTIALFÜHRUNG	OFFIS - INSTITUT FÜR INFORMATIK, OLDENBURG
KONSORTIALPARTNER	LANDESKREBSREGISTER NRW, BOCHUM
KOOOPERATIONSPARTNER	OFFIS CARE - KLINISCHE LANDESAUSWERTUNGSSTELLE / EPIDEMIOLOGISCHES KREBSREGISTER NIEDERSACHSEN, OLDENBURG

Inhalt

1 Präambel

Dieses Dokument beschreibt die Konzeption des Data Mining Werkzeugs, die im Rahmen des Projekts SePaMiM erstellt wurde. In Abschnitt 2 wird ein allgemeiner Überblick zum Data Mining im Kontext des SePaMiM-Projekts gegeben. In Abschnitt 3 werden verschiedene Techniken zum Clustering von sequentiellen Daten präsentiert. In Abschnitt 4 wird das geplante Vorgehen zur Erstellung des dazugehörigen Data Mining Werkzeugs skizziert.

2 Data Mining im Kontext von SePaMiM

Data Mining umfasst Techniken, um interessante Muster in großen Datenbeständen aufzudecken [1]. Data Mining Verfahren lassen sich in zwei grundlegende Kategorien unterteilen [1]: Überwachte und unüberwachte Verfahren. Überwachte Verfahren lernen anhand von zuvor gelabelten Daten ein Modell, mit dem Ergebnisse für neue, ungelabelte Daten vorhergesagt werden können. Beispiele für überwachte Verfahren sind Klassifikation und Regressionsanalysen. Unüberwachte Verfahren arbeiten auf ungelabelten Daten. Diese Verfahren dienen dazu, automatisch versteckte Muster in den Daten zu entdecken. Hierzu lernen die Verfahren selbstständig die Struktur der Daten, um Unterschiede oder interessante Eigenschaften eines Datensatzes, zum Beispiel anhand von Ähnlichkeitsmaßen, zu finden. Beispiele für unüberwachte Verfahren sind Clustering und Assoziationsanalysen.

Im SePaMiM Projekt sollen Data Mining Techniken genutzt werden, um in Behandlungsverläufen nach Mustern zu suchen. Das Konzept für das Data Mining Werkzeug basiert auf den vorab erarbeiteten Ergebnissen zum Stand der Technik bzgl. Data Mining auf Verlaufs- / Sequenzdaten bzw. Behandlungsverläufen. Viele Data Mining Aufgaben, wie Klassifikation, Regressionsanalysen, Clusteranalysen, etc., lassen sich auch auf sequentiellen Daten durchführen. Bei den hierzu verwendeten Verfahren handelt es sich teils um Erweiterungen klassischer Data Mining Verfahren aber auch um Neuentwicklungen [2].

Der Stand der Technik wurde dem LKR.NRW vorgestellt und fachlich diskutiert. Da es sich bei den Daten der klinischen Krebsregister um einen neuartigen und sehr komplexen Datensatz handelt, ist das Auffinden neuer Strukturen hier besonders relevant. Daher wurde zuerst beschlossen, den Fokus auf unüberwachte Verfahren zu legen. Hierdurch soll versucht werden, neue Muster in den Daten zu entdecken. Besonderes Interesse gab es bei der automatischen Aufdeckung von Patient*innen-Kohorten anhand ihrer Behandlungsverläufe. Hierbei erschien Clustering als besonders geeignet. In Folge dessen wurden noch einmal vertiefend die Eigenschaften von Clustering auf Verlaufsdaten untersucht und entschieden, den Fokus auf Clustering als Data Mining Verfahren zu setzen. Im folgenden Abschnitt wird daher ein Überblick über sequentielles Clustering gegeben.

3 Sequentielles Clustering

Das sequentielle Clustering befasst sich damit, eine Menge von Sequenzen anhand ihrer Ähnlichkeit in Gruppen einzuteilen [2]. Im Rahmen des SePaMiM Projektes soll sequentielles Clustering dazu genutzt werden, Patient*innenkohorten anhand von Behandlungsverläufe zu bilden. Verfahren für das Clustern von Sequenzen lassen sich in drei grundlegende Kategorien einteilen [3], [7]: ähnlichkeitsbasiertes sequentielles Clustering, indirektes sequentielles Clustering und statistisches sequentielles Clustering. Diese werden in den folgenden drei

Abschnitten vorgestellt. Da die Behandlungsverlaufsdaten im SePaMiM Projekt aus komplexen, multidimensionalen Ereignissen bestehen, werden im Anschluss Möglichkeiten zum Clustern solcher Daten beschrieben.

3.1 Ähnlichkeitsbasiertes sequentielles Clustering

Ähnlichkeitsbasiertes sequentielles Clustering nutzt bereits vorhandene distanzbasierte Clustering Verfahren, wie zum Beispiel hierarchische Verfahren, dichtebasierte Verfahren, wie DBScan und partitionierende Verfahren, wie K-Means [3], [7]. Die Herausforderung besteht hierbei darin, ein geeignetes Ähnlichkeitsmaß zu finden, das die Ähnlichkeit zweier Sequenzen bestimmt.

Ein einfacher Ansatz, die Ähnlichkeit zweier Sequenzen gleicher Länge zu bestimmen, besteht darin, gegenüberstehende Elemente der Sequenzen zu vergleichen und die einzelnen Ergebnisse zu aggregieren.

Sequenz-Alignment Ansätze, sind dagegen nicht auf Sequenzen gleicher Länge beschränkt. Diese versuchen, Sequenzen optimal aufeinander abzubilden. Hierdurch berücksichtigen sie außerdem zeitliche Verschiebungen innerhalb der Sequenzen. Hierunter fallen z.B. Editierdistanzen und Dynamic Time Warping (DTW).

Editierdistanzen werden insbesondere für Sequenzen bestehend aus Symbolen genutzt. Sie bestimmen die Ähnlichkeit zwischen zwei Sequenzen anhand der minimal benötigten Modifikationen (Einfügen, Löschen und Ersetzen), um eine Sequenz in die andere zu überführen [3]. Die Kosten für einzelne Modifikationen können dabei gewichtet werden, um zusätzlich Domänenwissen zu berücksichtigen.

DTW sucht die optimale Abbildung von zwei Sequenzen aufeinander. Hierbei werden jedoch keine Modifikationen vorgenommen, um eine Sequenz auf eine andere abzubilden. Stattdessen wird jedes Element einer Sequenz auf ein Element der Vergleichssequenz abgebildet. Es wird die Abbildung gesucht, die minimale Kosten verursacht, d.h. es werden möglichst ähnliche Elemente aufeinander abgebildet. Die Ähnlichkeit zwischen zwei Elementen kann dabei je nach Anwendungsfall unterschiedlich bestimmt werden. Bei der Abbildung wird außerdem eine Reihe von Einschränkungen berücksichtigt. Zum Beispiel kann nicht zurückgesprungen werden. Sobald ein Element der Eingangssequenz auf Element *n* der zu vergleichenden Sequenz abgebildet wurde, kann ein folgendes Element nicht mehr auf ein Element *n-1* oder kleiner der Vergleichssequenz abgebildet werden [6].

3.2 Indirektes sequentielles Clustering

Indirekte sequentielle Clustering Methoden verfolgen den Ansatz, Sequenzen zunächst in Feature-Vektoren umzuwandeln, um diese im Anschluss mit vorhandenen Clustering-Verfahren clustern zu können [3, [7]. Entsprechend wichtig ist hierbei die Auswahl geeigneter Features. Ein möglicher Ansatz besteht zum Beispiel darin, Sequenzen auf das Vorhandensein bzw. Nichtvorhandensein einzelner Elemente zu reduzieren. Alternativ kann auch die Häufigkeit der Elemente genutzt werden. Bei diesen Ansätzen geht jedoch die Reihenfolge der Elemente verloren. Eine Möglichkeit, die Reihenfolge der Elemente zu berücksichtigen, ist die Verwendung von kurzen Teilsequenzen. Eine Teilsequenz besteht aus *k* aufeinanderfolgenden

Elementen und wird als Feature genutzt, das angibt ob bzw. wie oft eine Teilsequenz vorhanden ist [7]. Bei numerischen Sequenzen können beispielsweise bekannte Verfahren aus der Signalverarbeitung, wie die Haar Wavelet Transformation genutzt werden, um die Sequenz in einen Feature-Vektor abzubilden [7], [10].

3.3 Statistisches sequentielles Clustering

Beim statistischen sequentiellen Clustering werden statistische Modelle für das Clustering genutzt. Hierbei wird die Annahme getroffen, dass jede Sequenz durch ein statistisches Modell bzw. durch Wahrscheinlichkeitsverteilungen erstellt wurde [7]. Die verschiedenen Ansätze lassen sich weiter anhand der Verwendung der statistischen Modelle in zwei Kategorien einteilen.

In der ersten Kategorie wird davon ausgegangen, dass die den Sequenzen zugrundeliegenden Modelle jeweils ein Cluster darstellen. Entsprechend werden hierbei die Parameter der Modelle erlernt, beispielsweise mit Hilfe des EM-Algorithmus (Expectation–maximization algorithm) [3]. Die zweite Kategorie beinhaltet Methoden, die statistische Modelle zur Berechnung der Ähnlichkeit zwischen Sequenzen nutzen [7]. Hierbei wird zunächst jede Sequenz auf ein Modell abgebildet und im Anschluss der Abstand zwischen zwei Sequenzen mittels Ähnlichkeitsmaßen für Wahrscheinlichkeitsverteilungen, wie z.B. der Kullback-Leibler-Divergenz ermittelt [7]. Zum Clustern werden dann auf Abstandsmaßen basierende Verfahren genutzt. In beiden Ansätzen werden u.A. Markov Ketten, Hidden Markov Modelle oder Autoregressive Moving Average (ARMA) Modelle als statistische Modelle genutzt [3], [7].

3.4 Umgang mit komplexen sequentiellen Daten

Viele Ansätze zum Clustern von sequentiellen Daten sind zunächst darauf ausgelegt, auf einfachen Sequenzen bestehend aus numerischen oder kategorialen Werten angewendet zu werden. Die Behandlungsverläufe, die im SePaMiM Projekt als Eingabe für das Clustering dienen, bestehen jedoch aus komplexen, multidimensionalen Ereignissen, die sowohl numerische als auch kategoriale Attribute besitzen. Zudem unterscheidet sich die Struktur der Ereignisse je Ereignistyp. Beispielsweise beinhalten Operationsereignisse u.a. eine Intention, eine Liste von Komplikationen, sowie eine Liste von OPS Codes. Strahlentherapien hingegen beinhalten u.a. ebenfalls eine Intention und zusätzlich eine Liste von Bestrahlungen und eine Liste von Nebenwirkungen.

Eine Möglichkeit, mit Sequenzen komplexer Daten umzugehen, besteht darin, die einzelnen Elemente zu kategorisieren und auf einfache Symbole abzubilden. Dieses Vorgehen wird zum Beispiel in [9] beschrieben. Die Autor*innen untersuchen die Versorgungspfade unterschiedlicher Patient*innen mit Herzinsuffizienz. Datengrundlage sind hierbei Abrechnungsdaten von Krankenkassen aus einer Periode von 24 Monaten. Aus diesen Daten werden einfache, aus Symbolen bestehende Sequenzen gebildet. Da die Abrechnungsdaten lediglich quartalsweise erhoben wurden, repräsentiert jedes Symbol einer Sequenz ein Abrechnungsquartal. Für jeden der drei Aspekte *Spezialisierung des Arztes*, *Prozedur* und *Medikation*, wird jeweils ein Sequenztyp gebildet. So wird beispielsweise für jede untersuchte Person eine Sequenz mit der Reihenfolge der Medikation erstellt, deren Symbole entweder die Ausprägung *ACE Hemmer* oder *Betablocker* haben. Die einzelnen Sequenztypen werden

dann individuell geclustert. Hierzu wird das k-Medoids Clustering Verfahren mit der Editierdistanz *Longest Common Subsequence* (LCS) genutzt.

Insbesondere im Bereich der auf Abstandsmaßen basierenden Clusteringverfahren konnten bereits verschiedene Beispiele gefunden werden, die auf komplexeren Daten ausgeführt werden. In [6] wird ein Clustering von Schneeprofilen beschrieben. Ein Schneeprofil besteht aus mehreren Schichten, die sich über die Zeit entwickeln. Jede Schneeschicht enthält Attribute, sowohl numerische als auch kategoriale, wie beispielsweise Schneekörnung und Schneehärte. Es werden verschiedene Abstandsmaße vorgestellt, um die einzelnen Attribute der Schneeschichten zu vergleichen. Hierauf aufbauend wird ein auf Dynamic Time Warping basierendes Clustering vorgestellt, bei dem die Attribute außerdem unterschiedlich gewichtet werden können. In [8] befassen sich die Autoren mit dem Clustern von multidimensionalen, semantischen Sequenzen. Als Anwendungsbeispiel dienen hier Tagesausflüge, die anhand der unternommenen Aktivitäten in Gruppen eingeteilt werden sollen. Eine Aktivität wird durch drei Attribute beschrieben: besuchter Ort, Art der Unternehmung und architektonischer Stil. Jedes Attribut ist ein Element einer Ontologie. Deren hierarchische Struktur wird zur Bestimmung des Abstandes von Attributen aus zwei unterschiedlichen Sequenzen herangezogen. Beispielsweise ähnelt der Besuch einer Bar dem Besuch eines Restaurants, da beide in der Ontologie als Orte, an denen etwas gegessen werden kann, aufgeführt werden. Zum Vergleich von zwei Sequenzen wird eine Editierdistanz verwendet. Die Autoren untersuchen unterschiedliche Clusteringverfahren und kombinieren diese zusätzlich mit dem Dimensionsreduktionsverfahren UMAP (Uniform manifold approximation and projection for dimension reduction).

4 Konzept Data Mining Werkzeug

Wie im vorherigen Abschnitt dargestellt, gibt es eine Vielzahl von vielversprechenden Ansätzen, um Sequenzen zu clustern. Im SePaMiM Projekt sollen unterschiedliche Ansätze zum Clustering verfolgt werden. Die Datengrundlage ist hierbei das in Zusammenarbeit mit dem LKR.NRW entwickelte Analysemodell, aus dem die für die jeweilige Analyse benötigten Daten exportiert und in ein geeignetes Format transformiert werden. Die Entwicklung des Data Mining Werkzeugs erfolgt iterativ mit regelmäßigen Feedbackrunden zwischen OFFIS und dem LKR.NRW.

Da sich Krebsarten im Verhalten erheblich untereinander unterscheiden können und daher verschiedene Behandlungsmaßnahmen erfordern, ist das Clustering anhand von Behandlungsverläufen insbesondere innerhalb einzelner Krebsentitäten sinnvoll. Daher wird der Fokus auf das Clustern von Brustkrebspatien*innen gelegt. Es wurde beschlossen, sich auf ähnlichkeitsbasierte sequentielle Clusteringverfahren zu fokussieren, da diese es besonders einfach machen, Expertenwissen in die Berechnung der Ähnlichkeit und somit in das Ergebnis des Clustering einfließen zu lassen. Die Ergebnisse des Clustering werden dann mit dem LKR.NRW diskutiert. Die hierbei gewonnenen Erkenntnisse sollen dann in die weitere Entwicklung des Data Mining Werkzeuges einfließen. Es ist daher davon auszugehen, dass hierdurch Anpassungen an dem im Folgenden skizzierten Konzept notwendig sein werden.

Ein erster Ansatz, der untersucht werden soll, ist die Kategorisierung von Ereignissen auf Symbole. Ein sehr einfacher Ansatz hierfür ist die Abbildung der Ereignisse des Analysemodells

auf ihren Ereignistyp, wie z.B. Operation oder Strahlentherapie. Dabei gehen jedoch viele Informationen verloren. Deshalb sollen die Sequenzen auf analyse-spezifisch abgeleitete Symbole abgebildet werden. Beispielsweise kann eine Operation im Kontext einer Brustkrebserkrankung anhand ihres OPS-Codes auf die Symbole BET (Brusterhaltene Therapie), Mastektomie bzw. sonstige Operation abgebildet werden. Ähnliche Abbildungen sind auch für andere Ereignistypen bzw. Krebsentitäten denkbar und werden im Verlauf der Entwicklung mit dem LKR.NRW besprochen und umgesetzt. Diese Sequenzen von Symbolen können dann im Anschluss geclustert werden. Als Abstandsmaße können hierzu beispielsweise Editierdistanzen genutzt werden. Ein solches Vorgehen soll unter Verwendung der Levenshtein Editierdistanz und einem hierarchischen Clusteringverfahren als erster Proof-of-Concept umgesetzt.

In einem weiteren Ansatz soll ein Ähnlichkeitsmaß entwickelt werden, das sowohl die Ähnlichkeit einzelner Ereignisse als auch die Reihenfolge der Ereignisse berücksichtigt. Dazu sollen Operationen, Strahlentherapien, systemischen Therapien und Krebsdiagnosen eines Patienten herangezogen werden. Die Ähnlichkeit zwischen verschiedenen Behandlungsereignissen soll anhand ausgewählter Attribute bestimmt werden. Hierzu wurden bereits erste Vorarbeiten vom LKR.NRW durchgeführt. Bei Operationen sollen zum Beispiel der OPS-Code, der Residualstatus und die Komplikationen zur Bestimmung des Abstandes herangezogen werden. Die berechneten Abstände zwischen einzelnen Ereignissen sollen dann als Kosten einer Levenshtein-Distanz genutzt werden.

Das Ergebnis des Clusterings ist die Zuordnung einzelner Patient*innen zu einer Kohorte anhand ihres Behandlungsverlaufs. Diese Ergebnisse sollen durch Domänenexpert*innen ausgewertet und interpretiert werden können. Neben der Möglichkeit, einzelne Patient*innen aus den Clustern zu betrachten, was insbesondere bei größeren Clustern sehr aufwändig werden kann, ist es hierzu sinnvoll eine Zusammenfassung der einzelnen Cluster zu erstellen. Zum einen können die Cluster über Verfahren der deskriptiven Statistiken beschrieben werden. Beispiele hierfür sind die Größe des Clusters oder die Geschlechtsverteilung innerhalb eines Clusters. Eine weitere Möglichkeit ist es, jedes Cluster durch einen repräsentativen Behandlungsverlauf darzustellen, beispielsweise durch den medoiden Behandlungsverlauf. Ebenso ist es denkbar, die Überlebenszeit der einzelnen Cluster zu berechnen, um die Cluster über diese Metrik miteinander zu vergleichen. Auch die Verwendung weiterer Data Mining Verfahren wie Sequential Pattern Mining ist eine Option. Hierbei werden in einer Menge von Sequenzen häufig auftretende Teilsequenzen gesucht [11]. Hiermit könnten häufig auftretende Teilbehandlungsverläufe eines Clusters ermittelt werden.

Literatur

1. Han J, Kamber M, Pei J (2011) Data Mining: Concepts and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
2. Laxman S, Sastry P (2006) A Survey of Temporal Data Mining. Sadhana 31:173–198
3. Xu R, Wunsch D (2005) Survey of Clustering Algorithms. Neural Networks, IEEE Transactions on 16:645–678

4. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26:43–49. https://doi.org/10.1109/TASSP.1978.1163055
5. Xing Zz, Pei J, Keogh E (2010) A Brief Survey on Sequence Classification. SIGKDD Explorations 12:40–48. https://doi.org/10.1145/1882471.1882478
6. Herla F, Horton S, Mair P, Haegeli P (2021) Snow profile alignment and similarity assessment for aggregating, clustering, and evaluating snowpack model output for avalanche forecasting. Geoscientific Model Development 14:239–258
7. Warren Liao T (2005) Clustering of Time Series Data-a Survey. Pattern Recogn. 38:1857–1874
8. Moreau C, Chanson A, Peralta V, De Runz C (2021) Clustering sequences of multi-dimensional sets of semantic elements. pp 384–391
9. Vogt V, Scholz S, Sundmacher L (2018) Applying sequence clustering techniques to explore practice-based ambulatory care pathways in insurance claims data. The European Journal of Public Health 28:214–219
10. Rani S, Sikka G (2012) Recent Techniques of Clustering of Time Series Data: A Survey. International Journal of Computer Applications 52:1–9. https://doi.org/10.5120/8282-1278
11. Fournier-Viger P, Chun J, Lin W, Kiran RU, Koh YS, Thomas R. A Survey of Sequential Pattern Mining. Vol. 1. 2017.