Aus dem Institut für Medizinische Informatik der Universität Heidelberg
Direktor: Prof. Dr. Martin Dugas

ÄHNLICHKEITSANALYSE TEMPORALER PATIENTENNETZWERKE AUF BASIS VON GRAFENBASIERTEN VERGLEICHSALGORITHMEN

Inauguraldissertation

zur Erlangung des Doctor scientiarum humanarum

an der Medizinischen Fakultät Heidelberg

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

JENS SCHRODT

aus Mühlacker

2023

Dekan: Prof. Dr. med. Dr. h.c. Hans-Georg Kräusslich

Doktormutter: Frau Prof. Dr. sc. hum. Petra Knaup-Gregori

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungen

API	Engl.: Application Programming Interface Dt.: Programmierschnittstelle
CBR	Case-based Reasoning
DBMS	Datenbank-Management-System
DDL	Data Definition Language
DML	Data Manipulation Language
EHR	Electronic Health Record
EIG	Eigenvalues Extraction
EMR	Electronic Medical Record
EUS	Entscheidungsunterstützungssystem
FSM	Frequent Subgraph Mining
GED	Graph Edit Distance
GUI	Engl.: Graphical User Interface Dt.: Grafische Benutzeroberfläche
JSON	JavaScript Object Notation
qSOFA	quick Sequential Organ Failure Assessment
SIRS	Systemic inflammatory response syndrome
SOFA	Sequential Organ Failure Assessment
SQL	Structured Query Language
XML	Extensible Markup Language

1 Einleitung

1.1 Gegenstand und Motivation

Entscheidungsunterstützungssysteme (EUS) sind rechnerbasierte Anwendungssysteme, welche menschlichen Entscheidungsträgern in wenig strukturierten oder komplexen Situationen bei der Entscheidungsfindung behilflich sein sollen. Im klinischen Bereich sollen sie dabei helfen, die Früherkennung, Diagnostik und Therapieentscheidung für Patienten zu verbessern. Betrachtet man zum Beispiel das Problem der im Krankenhaus erworbenen Infektionen durch multiresistente Keime, ist es wichtig, die Gefahr einer Sepsis frühzeitig zu erkennen, um lebensgefährdende Verläufe zu vermeiden. Die differenzierten Erscheinungsformen einer Sepsis sowie das Vorliegen verschiedener Kombinationen der im Kriterienkatalog der Deutschen Sepsis-Gesellschaft festgehaltenen Symptome, machen die rechtzeitige Erkennung einer Sepsis zu einer komplexen Aufgabe. Daher können EUS durch Einbeziehung sämtlicher relevanter Faktoren helfen, eine Sepsis frühzeitig zu diagnostizieren und eine adäquate Therapie einzuleiten.

Um ein EUS technisch umzusetzen, versuchen aktuelle Ansätze der Medizinischen Informatik automatisiert auf die Gesamtheit von Erfahrungen, die in der Vergangenheit bei der Behandlung von Patienten gemacht wurden, zurückzugreifen und darauf aufbauend Entscheidungen nachvollziehbar vorzuschlagen. Ein solches Verfahren stellt das fallbasierte Schließen (engl. Case-based Reasoning, CBR) dar. Hierbei werden alle historischen Behandlungsfälle durch eine strukturierte Menge an klinischen Parametern sowie der Therapie und des Behandlungsergebnisses beschrieben und in einer Fallbasis gespeichert. Für einen neuen Fall werden nach Möglichkeit die gleichen Parameter wie bei den historischen Fällen in der Fallbasis erfasst. Mithilfe der Berechnung eines oder mehrerer Ähnlichkeitsmaße wird versucht, die Ähnlichkeit zwischen je zwei Fällen zu quantifizieren. Somit können diejenigen Fälle aus der Fallbasis identifiziert werden, die dem aktuellen Fall bezüglich der klinischen Parameter am ähnlichsten sind, und die früheren Erfahrungen mit dem Behandlungsverlauf können in die Therapieentscheidung mit einbezogen werden.

Das fallbasierte Schließen greift das menschliche Vorgehen bei der Entscheidungsfindung auf: Auch ein Arzt stellt seine Diagnose und trifft seine Therapieentscheidung auf Basis der Vorgeschichte des Patienten. Darin eingeschlossen sind aufgetretene Symptome, die für die Erkrankung charakteristisch sind. Dabei greift der Arzt auf seine langjährige Erfahrung zurück und erkennt Krankheitsbilder oft gerade deshalb, weil er ein ähnliches Krankheitsbild in seiner Laufbahn schon einmal gesehen hat.

Verfahren, wie das fallbasierte Schließen, gewinnen durch die Verfügbarkeit immer größerer Datenmengen in der Medizin weiter an Bedeutung. Beispielsweise wird die Berücksichtigung der Ähnlichkeit von Patienten als wichtiger Schritt für die Präzisionsmedizin betrachtet (Brown 2016).

Ein weitergehender Aspekt bei der Betrachtung der Ähnlichkeit von Patienten sind die zeitlichen Beziehungen zwischen Ereignissen im Krankheitsverlauf sowie deren Ähnlichkeit bei verschiedenen Patienten. Für das oben genannte Beispiel der Sepsis wurde in verschiedenen Arbeiten der zeitliche Verlauf der Sepsis bei verschiedenen Patienten untersucht (Dreiher et al. 2012; Clermont et al. 2004; Wong et al. 2014). Auch in der deutschen Sepsis-Leitlinie werden für die Diagnose und Therapieentscheidungen zeitliche Einordnungen vorgenommen. Allerdings werden zeitliche Bezüge derzeit nur selten in klinischen EUS berücksichtigt. Da in modernen klinischen Informationssystemen Eintragungen in Patientenakten in der Regel mit Zeitstempeln versehen werden, könnten diese Informationen genutzt werden, um EUS um temporale Aspekte zu erweitern. Bisher gibt es kaum Methoden und Werkzeuge, die eine effiziente Nutzung dieser Daten für eine klinische Entscheidungsunterstützung ermöglichen. Für ein fallbasiertes Schließen wäre es beispielsweise notwendig, neue geeignete Darstellungsformen für temporale Daten sowie neue Ähnlichkeitsmaße zu entwickeln.

1.2 Problemstellung

Daten zu Behandlungsverläufen von Patienten liegen in der Regel in Form von Tabellen vor. Dabei werden temporale Zusammenhänge nicht explizit repräsentiert, sondern können über Zeitangaben von Ergebnissen rekonstruiert

werden. Anders ist es bei der Darstellung von Verläufen in der Form von Grafen. Hier werden temporale Zusammenhänge explizit repräsentiert und bereits existierende Verfahren der Grafentheorie können genutzt werden (Berlingerio et al. 2012). Dennoch gibt es zum Thema der temporalen Grafen bisher nur wenige Vorarbeiten. Die meisten beziehen sich auf eine lineare Darstellung zeitlicher Ereignisse, sogenannte Zeitstrahlen (Chen et al. 2017). Der Zeitstrahl ist als Darstellungsform für die Komplexität medizinischer Sachverhalte weitgehend unzureichend. Welche konkreten Eigenschaften passende temporale Grafen haben sollten, ist allerdings noch unklar und muss untersucht werden. Konkrete Eigenschaften wie ein passender temporaler Graf aussehen könnte, betreffen beispielsweise die Frage, in welcher Granularität die patientenbezogenen Ereignisse in Knoten überführt werden (temporale Abstraktion (Moskovitch und Shahar 2009)) und ob eine Typisierung der Knoten sinnvoll ist. Weiterhin ist unklar, welche Topologie der Grafen geeignet ist. Solche Topologien können beispielsweise eine Baumdarstellung oder das Erlauben von Zyklen für wiederkehrende Ereignisse sein. Von den für ein fallbasiertes Schließen zu verwendenden Grafen zur Darstellung temporaler Zusammenhänge sind die oft in der Literatur anzutreffenden temporalen Netzwerke abzugrenzen. Diese Netzwerke stellen keine temporalen Zusammenhänge dar, sondern sind nur in dem Sinne temporal, als dass die Entwicklung der Netzwerke über die Zeit festgehalten wurde und sich damit Rückschlüsse auf bestimmte Zusammenhänge ziehen lassen (Holme 2015). Auch diese Darstellungsform ist deshalb für den hier beschriebenen Zusammenhang unzureichend.

Für die Nutzung temporaler Grafen für eine klinische Entscheidungsunterstützung mit fallbasiertem Schließen ist darüber hinaus bisher kein Ähnlichkeitsmaß bekannt. Es gilt daher zu untersuchen, welche allgemeinen Verfahren der Grafentheorie zur Ermittlung der Ähnlichkeit zwischen Grafen für die Anwendung auf temporale Grafen geeignet sein könnten und somit als Grundlage für die Verwendung in Entscheidungsunterstützungssystemen genutzt werden können.

1.3 Zielsetzung

Aus der genannten Problemstellung ergeben sich folgende Ziele für die vorliegende Arbeit:

Ziel 1:

Erarbeiten einer grafenbasierten konzeptionellen Darstellungsform für temporale Ereignisse im Krankheitsverlauf von Patienten und deren technische Umsetzung zur formalen Repräsentation von klinischen Daten in einer Fallbasis für klinische Entscheidungsunterstützung.

Ziel 2:

Etablierung eines effizienten Speicherverfahrens für große Mengen an temporalen, grafenbasierten Krankheitsverläufen verschiedener Patienten unter Verwendung eines grafenfreundlichen Datenbank-Systems.

Ziel 3:

Identifizierung eines Ähnlichkeitsmaßes zur Nutzung von grafenbasierten temporalen Daten in einem CBR-System mit einer Fallbasis, die in einem grafenorientierten oder anderem geeigneten Datenbank-System gespeichert ist.

1.4 Vorgehensweise

Grundlage für die Entwicklung der Elemente eines grafenorientierten CBR-Systems bildet der für Forschungszwecke frei verfügbare Datensatz MIMIC III mit ca. 40.000 Krankheitsverläufen einer Intensivstation (Johnson et al. 2016). In dieser Datenbank werden typische Fälle identifiziert und zur Entwicklung der Methodik für die grafenorientierte Darstellung herangezogen. Weiterhin wird der Datensatz für die Untersuchung grafenorientierter Datenbank-Systeme sowie die Erprobung des Ähnlichkeitsmaßes genutzt. Ausgangspunkt für die Identifizierung eines geeigneten Ähnlichkeitsmaßes sind bekannte Methoden aus der Grafentheorie für die Quantifizierung der Ähnlichkeit zwischen Grafen. In einem ersten Schritt wird allerdings eine systematische Literaturstudie durchgeführt, mit dem Zweck, diejenigen Veröffentlichungen zu identifizieren,

die thematisch in dieselbe Richtung gehen und somit eine Einbettung der vorliegenden Arbeit in das aktuelle Forschungsfeld zu schaffen.

2 Grundlagen

2.1 Sepsis

Unter den Todesursachen in Deutschland ist die Sepsis nach wie vor unter den zehn häufigsten und kann ungefähr auf eine Stufe mit der Anzahl der Todesursachen durch Herzinfarkt gestellt werden. Im Jahr 2013 gab es 279.530 Fälle von Sepsis in Deutschland, diese endeten in 67.849 Fällen tödlich. 2017 kam eine IHME-Studie nach wie vor auf geschätzt 279.000 Fälle von Sepsis pro Jahr nach der neuen Sepsis-Definition (Fleischmann-Struzek et al. 2022). Je nach Fortschreiten der Sepsis liegt die Mortalitätsrate dabei zwischen 25-50% (Fleischmann et al. 2016).

Epidemiologie

Sepsis ist eine weltweit verbreitete Krankheit, die in erster Linie in Krankenhäusern auftritt. Die Letalitätsrate liegt bei 25-50%. In den USA stieg die Inzidenz der Sepsis jährlich zwischen 2000 und 2008 um 7-8% von 221/100.000 Einwohner (im Jahr 2000) auf 377/100.000 Einwohner (im Jahr 2008). 2007 starben in den USA mehr als 200.000 Patienten an einer Sepsis. Mögliche Gründe für den Inzidenzanstieg sind der demografischer Wandel sowie eine Zunahme medikamentöser, invasiver und immunsuppressiver medizinischer und intensivmedizinischer Maßnahmen (Fleischmann et al. 2016).

Auch in Deutschland stieg die Anzahl der Fälle an und zwar um ca. 5,7% von 200.535 (2007) auf 279.530 (2013). Das ergibt einen Inzidenzanstieg von 256/100.000 Einwohner auf 335/100.000 Einwohner. Die Sterbefälle lagen dabei 2013 bei 67.849, was eine Sterblichkeitsrate von 24,3% ergibt (Fleischmann et al. 2016).

Pathogenese

Eine Sepsis kann aus unterschiedlichen Infektionen entstehen. Auslöser einer solchen Infektion können Bakterien, Viren, Pilze und Parasiten sein. Definiert war die Sepsis bis 2016 nach den Kriterien des systemisch inflammatorischen Response-Syndroms (SIRS) (s.

Tabelle 1). Dabei war eine Sepsis immer dann zu diagnostizieren, wenn zwei oder mehr der SIRS Kriterien zutreffen und SIRS aus einer Infektion resultiert (Stearns-Kurosawa et al. 2011; Levy et al. 2003). In den weitaus meisten Fällen liegt der Infektionsherd dabei in der Lunge, im Abdomen oder im Harntrakt. Die Kriterien nach SIRS sind entweder Fieber, das heißt eine Temperatur über 38°C, oder eine Hypothermie, das heißt eine Temperatur kleiner oder gleich 36°C. Ein weiteres Kriterium für SIRS ist eine Tachykardie mit einer Herzfrequenz von größer oder gleich 90 Herzschlägen pro Minute. Des Weiteren gelten eine Tachypnoe (Atemfrequenz größer oder gleich 20 Atemzüge pro Minute) oder eine Hyperventilation mit einem PaCO2-Wert kleiner oder gleich 4,3 kPa bzw. 33mmHg ebenfalls als Kriterien nach SIRS. Die letzten Kriterien beziehen sich auf die Anzahl weißer Blutzellen, demnach gelten eine Leukozytose (mehr als 12.000 Leukozyten pro mm^3 Blut) oder eine Leukopenie (weniger als 4.000 Leukozyten pro mm^3 Blut) oder mehr als 10% unreife Neutrophile im Differentialblutbild als Kriterien, die eine Sepsis mit definieren.

Diese Definition der Sepsis (SEPSIS-2) wurde im Februar 2016 durch die SEPSIS-3-Definition ersetzt, wonach eine Sepsis als „lebensbedrohliche Organdysfunktion, verursacht durch eine fehlgeleitete Wirtsantwort auf eine Infektion“ definiert wird (Schmoch et al. 2017; Herold 2017). Dabei wird als septischer Schock diejenige Sepsis-Untergruppe definiert, in der „zugrundeliegende zirkulatorische, zelluläre und metabolische Anomalitäten mit signifikant größerem Letalitätsrisiko assoziiert sind als bei einer Sepsis alleine“ (Herold 2017). Der Begriff SIRS wird nach der neuen SEPSIS-3-Definition nicht mehr verwendet.

Tabelle 1: Sepsiskriterien der SEPSIS-2-Definition nach Levy et al. 2003), Stearns-Kurosawa et al. 2011), (Herold 2017).

Kriterium	Wert
Temperatur	Fieber (≥38,0°C) oder Hypothermie (≤36,0°C) bestätigt durch eine rektale, intravasale oder intravesikale Messung
Herzfrequenz	Tachykardie mit Herzfrequenz ≥90/min
Respirationsrate	Tachypnoe (Frequenz größer oder gleich 20/min) oder Hyperventilation (bestätigt durch Abnahme einer arteriellen Blutgasanalyse mit PaCO2 ≤4,3kPa bzw. 33mmHg)
Anzahl weißer Blutzellen	Leukozytose (≥12.000/mm³) oder Leukopenie (≤4.000/mm³) oder 10% oder mehr unreife Neutrophile im Differentialblutbild

Fieber wurde als eines der wichtigsten Symptome der Sepsis identifiziert (Assmann et al. 1949). Tabelle 1 zeigt aber auch, dass für die Identifikation einer Sepsis nicht zwangsläufig Fieber als Symptom vorhanden sein muss, es kann genauso auch eine Hypothermie oder keine von der Normaltemperatur abweichende Temperatur vorliegen und trotzdem kann die Diagnose laut SEPSIS-2-Definition eine Sepsis sein. Werdan et al. 2016 stellen außerdem klar, dass zwar eine Sepsis auf Basis der klinischen Kriterien wahrscheinlich ist, in 30% der Fälle allerdings kein mikrobiologisch gesicherter Infektionsnachweis geführt werden kann. Diese Beispiele machen deutlich, dass die Diagnose Sepsis durch Vielzahl der im Katalog beschriebenen Symptome und die unterschiedlichen Ausprägungen der Symptome bei den Patienten geprägt ist und somit schwierig zu diagnostizieren ist.

Die SEPSIS-3-Definition beinhaltet ebenso einen Kriterienkatalog (s. Tabelle 2), nach dem eine Sepsis vorliegt, wenn ein Verdacht auf eine Infektion und eine lebensbedrohliche Organdysfunktion besteht. Diese Organdysfunktion wird anhand der in Tabelle 2 verwendeten Werte des SOFA-Scores (Sequential Organ Failure Assessment Score) bewertet. Wenn mindestens zwei Punkte im SOFA-Score erreicht sind (s. Tabelle 2), dann liegt eine Sepsis vor. In den Zeilen der Tabelle 2 werden dabei die einzelnen Organe betrachtet, in den Spalten die jeweiligen Werte, die über- oder unterschritten werden müssen. Werden die jeweiligen Bedingungen erfüllt, dann werden die Punkte in den Spaltenköpfen vergeben. Ein septischer Schock liegt dagegen dann vor, wenn die Bedingungen für eine Sepsis erfüllt sind, also mindestens zwei Punkte im SOFA-Score erreicht sind, und zusätzlich die folgenden Bedingungen erfüllt sind: Serum Laktat > 2 mmol/l sowie Vasopressor-abhängige Hypotension trotz adäquater Flüssigkeitssubstitution (um MAP $\geq$ 65 mmHg zu halten) (Herold 2017).

Tabelle 2: Diagnostischer Kriterienkatalog nach der SEPSIS-3-Definition für Sepsis und septischen Schock entnommen aus Herold (2017).

SOFA-Score-Punkte	**1**	**2**	**3**	**4**
Lunge: PaO_2/FiO_2, mmHg	< 400	< 300	< 200 mit maschineller Beatmung	< 100 mit maschineller Beatmung
Gerinnung: Thrombozyten x $10^3/mm^3$	< 150	< 100	< 50	< 20
Leber: Bilirubin, mg/dl µmol/l	1,2-1,9 20-32	2,0-5,9 / > 4 33-101 / >70	6,0-11,9 102-204	> 12,0 > 204
Herz/Kreislauf: Hypotension	MAP < 70 mmHg	Dopamin < 5 oder Dobutamin (jede Dosierung)	Dopamin 5,1-15 oder Adrenalin ≤ 0,1 oder Noradrenalin ≤ 0,1 *)	Dopamin > 15 oder Adrenalin > 0,1 oder Noradrenalin ≤ 0,1 *)
ZNS: Glasgow Coma Scale	13-14	10-12	6-9	> 6
Niere: Kretinin, mg/dl µmol oder Diurese	1,2-1,9 110-170	2,0-3,4 171-299	3,5-4,9 300-400 oder < 500 ml/d	> 5,0 > 440 Oder < 200ml/d

*) Adrenerge Substanzen für mindestens 1h (Dosierung in µg/kg x Min)

Die häufigsten Erreger der Sepsis sind mit 55 % gramnegative und grampositive Bakterien, zu 20% sind es Pilze. Die häufigsten gramnegativen Erreger sind *Escherichia coli*, *Klebsiellae*, *Pseudomonas aeruginosa*, die häufigsten grampositiven sind *Staphylococcus aureus* und *Streptococcus pneumoniae* (Classen et al. 2010).

Therapie

Je nach Schweregrad der Sepsis - Systemisches inflammatorisches Response-Syndrom (SIRS), Sepsis, schwere Sepsis und septischer Schock in aufsteigender Reihenfolge nach SEPSIS-2-Definition - wird bei hohem Schweregrad zunächst eine ausreichende hohe Sauerstoffversorgung durch Beatmung und ein ausreichend hoher Blutdruck durch Flüssigkeitszugabe per Infusion sichergestellt. Ist der Blutdruck dennoch zu niedrig oder die Organdurchblutung nicht gewährleistet, so werden Katecholamine verabreicht. Anschließend, oder bei niedrigem Schweregrad der Sepsis, werden die Erreger im Infektionsherd bekämpft. Bei einer bakteriellen Infektion werden dabei Antibiotika verabreicht, bei einer Pilzinfektion Antimykotika. Nur durch Bekämpfung der Erreger und Beseitigung des Infektionsherdes gibt es langfristig auch Aussicht auf Heilung. Zusätzliche Therapiemaßnahmen sind Dialyse bzw. Hämofiltration und Ersatz von Blutzellen. Es werden außerdem drei Therapieformen unterschieden: die Early Goal Directed Therapy, die Standardtherapie und die adjuvante Therapie (Classen et al. 2010).

2.2 Grafen

In dieser Arbeit werden Grafen bzw. Netzwerke (die beiden Begriffe werden in dieser Arbeit gleichbedeutend verwendet) zur Vergleichsanalyse von Patientendaten verwendet. Die Theorie komplexer Netzwerke bzw. die Grafentheorie spielt eine große Rolle in recht unterschiedlichen wissenschaftlichen Disziplinen, unter anderem in der Informatik, der Soziologie, der Physik und in der Populations- und Molekularbiologie (Pavlopoulos et al. 2011). In der Grafentheorie werden Grafen als Netzwerke von Punkten definiert,

sogenannte Knoten (engl.: Vertices / Nodes), die über Linien zwischen den Punkten, sogenannte Kanten (engl.: Edge), miteinander verbunden sind (Newman 2016). Die Knoten stellen dabei die Objekte von Interesse dar (engl.: Entity), während die Kanten deren Beziehungen untereinander repräsentieren. Kanten können dabei beschrieben werden durch Gewichtung, Richtung und Typ, wobei jede Kante immer nur zwei Nodes miteinander verbindet (Huber et al. 2007).

Die beschriebenen Zusammenhänge werden in Abbildung 1 veranschaulicht. Punkte in einem Grafen werden als Knoten (engl. Node) definiert, die Verbindungen zwischen den Knoten werden Kanten (engl. Edge) genannt. Dabei können Kanten gerichtet oder ungerichtet sein, wobei gerichtete Kanten durch einen Pfeil in die entsprechende Richtung dargestellt werden (s. Abbildung 1 (b)). Solche gerichteten Kanten können nur in Richtung des Pfeils durchlaufen werden. Es gibt auch Kanten, die in beide Richtungen zeigen und damit Knoten in beide Richtungen miteinander verbinden. Das heißt der erste Knoten ist über eine Kante mit dem zweiten Knoten verbunden und der zweite Knoten weist wiederum eine Verbindung durch eine andere Kante zum ersten Knoten auf (s. Abbildung 1 (c)). Abbildung 1 stellt nur einen Beispielgrafen zur Veranschaulichung der Begrifflichkeiten dar. Bei der Anwendung von Grafen sollten allerdings nur gerichtete oder ungerichtete Grafen und keine Mischung der beiden verwendet werden (Huber et al. 2007).

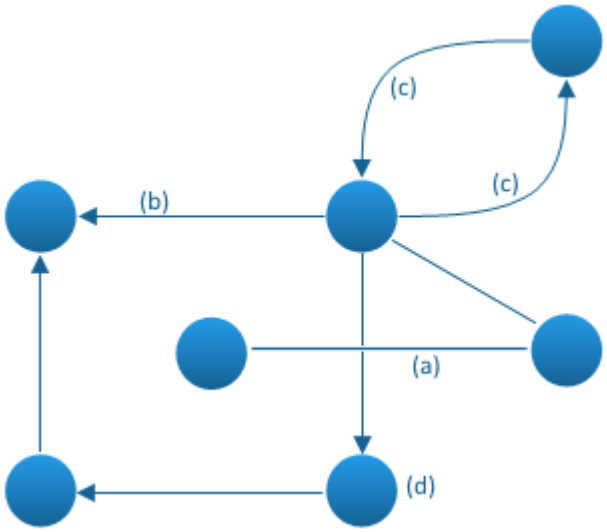

Abbildung 1: Schematische Darstellung eines Beispiel-Grafen zur Veranschaulichung der Begrifflichkeiten. (a) ungerichtete Kante (Edge). (b) gerichtete Kante, veranschaulicht durch einen Pfeil. (c) Edges, die von einem Node zum anderen führen und wieder zurück. (d) Node / Knoten.

Die Grafentheorie ist ein weit gefasstes Feld, welches auf bestimmten Regeln und Standards fundiert ist und deren Algorithmen und Methoden in den unterschiedlichsten Forschungsbereichen wie der Physik, Soziologie, Mathematik, Informatik, Populationsbiologie und Molekularbiologie eingesetzt werden (Pavlopoulos et al. 2011). Aus diesem Grund existieren im Bereich der Grafentheorie auch die unterschiedlichsten Algorithmen (Siek et al. 2002) und mathematischen Forschungsergebnisse, die aber alle auf denselben Begrifflichkeiten beruhen und deshalb auch für andere Zwecke angewendet werden können. Dabei „werden grafentheoretische Modelle und Algorithmen in unterschiedlichen Bereichen der Entscheidungsunterstützung eingesetzt. Dies liegt […] daran, dass vielfältige effiziente Algorithmen zur Lösung grafentheoretischer Problemstellungen entwickelt wurden. […]“ (Werners 2008). Die Grafentheorie stellt damit eine Art Abstraktionsebene für alle möglichen Anwendungen dar, wobei die Algorithmen unter den Anwendungen theoretisch ausgetauscht werden können.

2.3 Datenhaltung

Die heute typischen Datenbanken sind relationale Datenbanken. Dabei werden Datenbanktabellen durch Relationen miteinander in Beziehung gesetzt. Das Prinzip der Relationen beschreibt einen Verweis in einer Datenbanktabelle auf eine andere Datenbanktabelle bzw. Relation gemäß den Normalisierungsregeln, sodass dabei eine Form der Tabelle entsteht, in der keine - oder zumindest keine vermeidbaren - Redundanzen mehr auftreten (Codd 1991). Seit dem Jahr 2000 hält das Web 2.0 Einzug in das Internet und bringt damit einige Probleme für klassische relationale Datenbanken. Die relationalen Datenbanken sind in erster Linie auf den Ein-Server-Betrieb ausgelegt, was es schwierig macht, diese Datenbanken auf eine Weise zu skalieren, die effizient ist und trotzdem Daten im Bereich von Terabyte oder Petabyte verarbeiten kann. In diesem Zuge entstanden die NoSQL-Datenbanken, die darauf ausgelegt sind, die Skalierbarkeit des Datenbanksystems zu gewährleisten und so auch riesige Datenmengen zu verarbeiten. Eine Unterkategorie der NoSQL-Datenbanken sind die sogenannten Grafdatenbanken. Diese Datenbanksysteme benötigen

keine Tabellenschemata. Beispiele für solche Datenbanken sind db4o, neo4j, memcached und InfoGrid (Edlich et al. 2010). Im Folgenden soll etwas genauer auf die Grafdatenbank neo4j eingegangen werden. Neo4j wird mit einer eigenen Abfragesprache namens Cypher ausgeliefert, mit der vergleichsweise einfach Anfragen an die Datenbank gestellt werden können, die für die Algorithmen der Grafentheorie geeignet sind. Cypher wurde bei der Entwicklung außerdem an die Abfragesprache Structured Query Language (SQL) angelehnt (neo4j 2017). Ein Vorteil im Hinblick auf Performance der Grafdatenbanken gegenüber den relationalen Datenbanken ist außerdem, dass Grafdatenbanken die gemachten Verbindungen unter den Knoten speichern, während relationale Datenbanken ihre Relationen nicht speichern, diese müssen jedes Mal aus den zugehörigen Datensatznummern berechnet werden (neo4j 2017). Gerade bei komplexeren, hochverzweigten Strukturen mit vielen Beziehungen untereinander ist eine Grafdatenbank schneller als eine relationale Datenbank (Jaiswal und Agrawal 2015). Die Strukturierungsmöglichkeit der Daten in einer Grafdatenbank ist in dieser Arbeit im Vergleich zu relationalen Datenbanken außerdem von großem Vorteil.

2.4 MIMIC-III-Datenbank

Die MIMIC-III (Medical Information Mart for Intensive Care III) Critical Care Database (Saeed et al. 2002) ist eine frei verfügbare Datenbank der Intensivstation des Beth Israel Deaconess Medical Centers in Boston, Massachusetts, mit Daten von über 40.000 Patienten. Die Daten sind Echtdaten und müssen deshalb im Sinne des Datenschutzes mit besonderer Vorsicht genutzt werden. Aus diesem Grund wurde die Datenbank anonymisiert und konnte so für Forscher weltweit zur Nutzung zur Verfügung gestellt werden. Durch die Erfassung aller Ereignisse mit einem Zeitstempel ist diese Datenbank besonders gut geeignet für die Verwendung in dieser Arbeit, da der Zeitstempel Rückschlüsse auf temporale Zusammenhänge der einzelnen Ereignisse erlaubt. Dabei beinhaltet die Datenbank prinzipiell alle Daten, die in einer Intensivstation auch erfasst werden. Dazu gehören beispielsweise Vitalzeichen, Medikation, Laborwerte und Diagnosen (Johnson et al. 2016). Die Daten werden in dieser

Datenbank in relationaler Form gehalten und sind deshalb, anders als bei grafenbasierten Datenbanken, nicht direkt miteinander verbunden, sondern über IDs der jeweiligen Tabellenzeilen, also über Relationen.

2.5 Patientendaten als Grafen

Heutzutage sind elektronische Patientendatensätze (engl. Electronic Medical Record (EMR)) die vorherrschende Methode zur Erfassung und Dokumentation von Leistungen des Gesundheitswesens. Während erste Ansätze des EMR bereits vor mehr als zwei Jahrzehnten etabliert wurden, gibt es noch immer umfassende Forschungsaktivitäten rund um das Thema EMR auch im Zuge der Digitalisierung des Gesundheitswesens.

Aus informationstechnischer Sicht hat sich auch die Art und Weise der Speicherung von EMR-Daten über die Jahre verändert. Einer der ersten allumfassenden Ansätze war das Speichern von EMR-Datensätzen in relationalen Datenbanken (Friedman et al. 1990). Heutzutage basieren wahrscheinlich die meisten Krankenhausinformationssysteme auf relationalen Datenbanken Management Systemen und zugehörigen Datenabfragesprache SQL (Codd 1970; ISO ISO/IEC 9075 -1:2016).

Allerdings hat die Entwicklung hin zu NoSQL-Datenbankmanagementsystemen auch die Entwicklung von EMR-Systemen beeinflusst (Klein et al. 2015; Ercan und Lane 2014). So erübrigte sich bei dokumentenorientierten Datenbanksysteme wie CouchDB (Rascovsky et al. 2012) oder MongoDB (Luo et al. 2016; Xu et al. 2014) die Organisation der Daten in Tabellen. Stattdessen werden die Daten in Dokumenten in Datenformaten wie JavaScript Object Notation (JSON) (Bray 2017) oder der Extensible Markup Language (XML) (Bray et al. 2008) gespeichert. Ein weiterer Typ von NoSQL-Datenbanken sind Grafdatenbanken (Soulakis et al. 2015; Zhou et al. 2006). Im Gegensatz zu dokumentenbasierten Datenbanken, werden die Daten als Eigenschaften in Strukturen gespeichert, die aus Knoten und Kanten bestehen. Grafdatenbanken haben viele Vorteile gegenüber relationalen Datenbanken, beispielsweise sind Grafdatenbanken einfacher zu skalieren, sind schneller speziell bei

hochvernetzten Datensätzen und arbeiten auf einem höheren Niveau der Verfügbarkeit als gewöhnliche relationale Datenbanken (Nayak et al. 2013).

Grafdatenbanken (siehe auch Kapitel 2.3) werden bereits in verschiedenen sozialen Netzwerken wie Facebook und in anderen Internetfirmen wie Amazon oder Google genutzt (Moniruzzaman und Hossain 2013). In sozialen Netzwerken können Grafdatenbanken sehr nützlich sein, da sie die Beziehungen zwischen Teilnehmern des sozialen Netzwerkes direkt und intuitiv für den Anwender speichern können. Dieses direkte Speichern der Beziehungen der Teilnehmer reduziert die Rechenzeit und bietet die Möglichkeit Anfragen zu generieren, die diese Beziehungen direkt abfragen können. Diese gespeicherten Verbindungen zu nutzen, ist in Grafdatenbanken also sehr viel einfacher als die Nutzung von SQL-Anfragen im relationalen Datenbankmodell. Im relationalen Datenbankmodell wären komplexe join-Anfragen notwendig, um dieselben Effekte erzielen zu können. Das erhöht die Komplexität bei der Erstellung von Anfragen und erhöht ebenfalls die Rechenzeit im Vergleich zu Anfragen von Grafdatenbanken (Moniruzzaman und Hossain 2013). Außerdem gibt es noch viele weitere Anwendungsfelder für Grafdatenbanken neben den sozialen Netzwerken, beispielsweise biomolekulare Signalpfade (Fabregat et al. 2018), zur Vernetzung von heterogenen biologischen Daten (Yoon et al. 2017) und zur Darstellung von Netzwerken unterschiedlicher Krankheiten (Lysenko et al. 2016).

Grafen im Kontext der Grafentheorie werden klar definiert als ein Satz von Knoten, die durch Kanten miteinander verbunden sind. Diese Kanten stellen die Beziehung der Knoten miteinander dar (Bollobás 2010). Diese Definition des Begriffs Graf wird durchgehend in der gesamten Arbeit genutzt. Abbildung 2 zeigt die schematische Darstellung eines Grafen. Die Punkte stellen die Knoten eines Grafen dar, die Verbindungen zwischen den Knoten repräsentieren die Kanten.

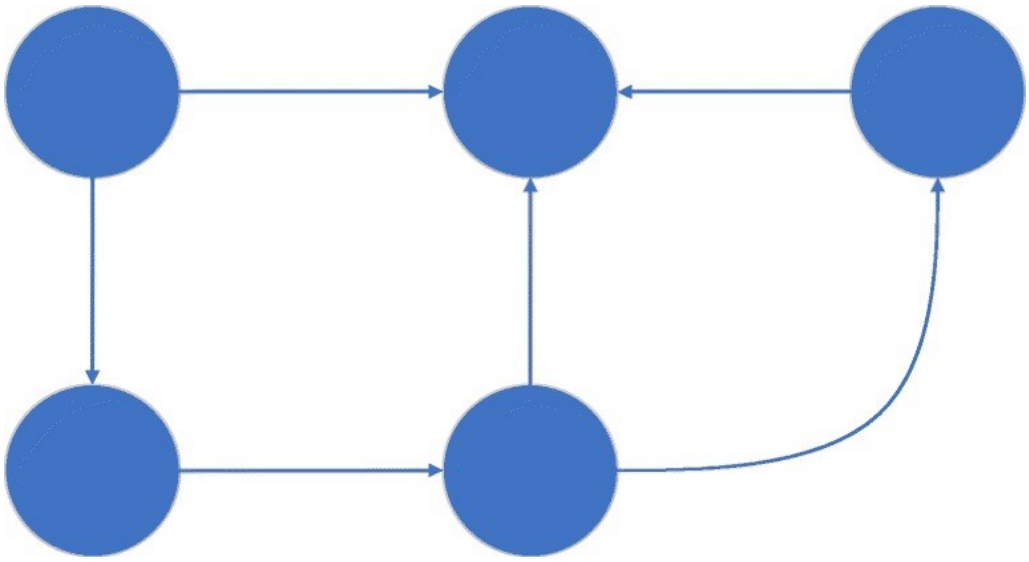

Abbildung 2: Schematische Darstellung eines gerichteten Grafen. Die Punkte werden Knoten genannt, die Verbindungen zwischen den Knoten werden Kanten genannt. Die Kanten sind gerichtet, das wird dargestellt durch die Pfeilrichtung, die jede Kante hat. Darstellung entnommen aus Schrodt et al. (2020).

Die Grafentheorie ist ein etabliertes Gebiet der Mathematik, das auch Methoden zum Vergleich von Grafen bereithält. Diese Methoden machen die Nutzung von Grafen im medizinischen Kontext sehr interessant, beispielsweise um die Patientendaten eines EMR-Systems als Graf zu modellieren. Mit Hilfe solcher Ansätze können Diagnosen, Therapien und Medikationen automatisiert vorgeschlagen werden, auf Basis von Daten vorheriger Patienten und der Erfahrung, die bei der Behandlung dieser Patienten gemacht wurde. Ein solches System könnte beispielsweise Teil eines Entscheidungsunterstützungssystems für Ärzte im klinischen Kontext sein. Beispielsweise werden Grafen dazu genutzt, um eine räumliche Beschreibung zerebraler Anatomie darzustellen (Atif et al. 2007) oder um Cluster von Patienten zu erstellen und Diagnosen zu stellen (Liu et al. 2015). Andere Ansätze sind näher verbunden mit EMR-Daten. Solche Projekte fokussieren sich beispielsweise auf die Visualisierung von EMR-Daten-Nutzung verbunden mit der Diagnose Herzinsuffizienz (Soulakis et al. 2015), auf die Modellierung von Krankheitsbildern (Yousefi et al. 2009) oder auf die Vorhersage unbekannter nachteiliger Medikament-Reaktionen (Bean et al. 2017).

2.6 NetSimile-Algorithmus

Ein Ziel dieser Arbeit ist die Identifizierung eines Ähnlichkeitsmaßes, welches dazu verwendet werden kann, Patientengrafen miteinander zu vergleichen. In diesem Kapitel wird deshalb der NetSimile Algorithmus vorgestellt, der diese Aufgabe erfüllen kann. Aufgabe des NetSimile-Algorithmus ist die Berechnung der Ähnlichkeit zweier Grafen unabhängig von der Größe der Grafen. Dazu wird ein Ähnlichkeitswert berechnet, der insbesondere die Struktur der Grafen beurteilt und daraus Rückschlüsse auf deren Ähnlichkeit zieht (Berlingerio et al. 2012). Tabelle 3 stellt den allgemeinen Ablauf des NetSimile-Algorithmus dar. Dabei werden zunächst bestimmte Eigenschaften (hier: „Features“) eines Grafen berechnet, diese werden anschließend in sogenannte „Signatur“-Vektoren aggregiert und im Anschluss daran werden die Signatur-Vektoren mit Hilfe der Canberra-Distanz miteinander verglichen. Je kleiner die Distanz, desto höher ist dabei die Ähnlichkeit (s. Tabelle 3).

Im Folgenden wird der Algorithmus genauer vorgestellt. Ein wichtiger Bestandteil ist das sogenannte Egonetz, das definiert wird als Netzwerk der direkten Nachbarn eines betrachteten Knotens. Im ersten Schritt des NetSimile-Algorithmus werden zunächst bestimmte Eigenschaften eines jeden Knoten in einem Grafen berechnet (s. Tabelle 4). Berlingerio et al. (2012) schlagen die folgenden Eigenschaften aufgrund ihrer Grafgrößenunabhängigkeit vor:

- Die Zahl der Nachbarknoten eines Knoten,
- den Clusterkoeffizienten eines Knoten,
- die durchschnittliche Anzahl an Knoten, die in der Nachbarschaft der Nachbarknoten liegen (würde man von Freunden einer Person ausgehen, die den zu untersuchenden Knoten darstellt, dann wäre die gesuchte Knotenzahl die durchschnittliche Zahl der Freundesfreunde pro Freund),
- den durchschnittlichen Clusterkoeffizienten der umgebenden Nachbarknoten eines Knoten,
- die Zahl der Kanten, die von dem zu untersuchenden Knoten ausgehen,
- die Zahl der Kanten, die von dem Egonetz des zu untersuchenden Knoten ausgehen sowie
- die tatsächliche Zahl der Nachbarsnachbarn (oder Freundesfreunde).

Diese Werte werden für jeden Knoten in dem untersuchten Grafen berechnet. Dadurch entsteht eine $Knoten\ x\ Feature$ Matrix für jeden Grafen. Diese Matrix wird im nächsten Schritt, der Aggregation, weiterverwendet.

Tabelle 3: Übersicht über den Algorithmus von NetSimile

Algorithmus 1 NETSIMILE

Require: $(\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\}), \boldsymbol{doClustering}$

1: // berechne Merkmale der einzelnen Knoten

2: $\{\boldsymbol{F_{G_1}}, \boldsymbol{F_{G_2}}, \dots, \boldsymbol{F_{G_k}}\}$:=GETFEATURES($\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\}$)

3: // erzeuge „signature“ Vektoren für jeden Grafen

4: $\{\vec{\boldsymbol{s}}_{\boldsymbol{G_1}}, \vec{\boldsymbol{s}}_{\boldsymbol{G_2}}, \dots, \vec{\boldsymbol{s}}_{\boldsymbol{G_k}}\}$:=AGGREGATOR($\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\}$)

5: // vergleiche und gib den Ähnlichkeits-/Abstands-Wert für die gegebenen Grafen zurück

6: **return** COMPARE($\{\vec{\boldsymbol{s}}_{\boldsymbol{G_1}}, \vec{\boldsymbol{s}}_{\boldsymbol{G_2}}, \dots, \vec{\boldsymbol{s}}_{\boldsymbol{G_k}}\}, \boldsymbol{doClustering}$)

Tabelle 4: Algorithmus der GetFeatures-Methode. Die Methode berechnet die Merkmale für die einzelnen Knoten und gibt sie in Knoten x Merkmal Matrizen zurück. $\boldsymbol{d_i} = |\boldsymbol{N(i)}|$: Anzahl der Nachbarn des Knoten $\boldsymbol{i}$. $\boldsymbol{c_i}$: Clusterkoeffizient des Knoten $\boldsymbol{i}$, $\overline{\boldsymbol{d}}_{\boldsymbol{N(i)}} = \frac{\boldsymbol{1}}{\boldsymbol{d_i}} \sum_{\forall \boldsymbol{j} \in \boldsymbol{N(i)}} \boldsymbol{c_j}$: Durchschnittliche Zahl an Nachbarsnachbarn, $\overline{\boldsymbol{c}}_{\boldsymbol{N(i)}} = \frac{\boldsymbol{1}}{\boldsymbol{d_i}} \sum_{\forall \boldsymbol{j} \in \boldsymbol{N(i)}} \boldsymbol{c_j}$: Durchschnittlicher Clusterkoeffizient über alle Nachbarn eines Knoten, $|\boldsymbol{E_{ego(i)}}|$: Anzahl der Kanten in direkter Nachbarschaft des Knoten. $|\boldsymbol{E^o_{ego(i)}}|$: Anzahl der Kanten, die von dem Knoten ausgehen, $|\boldsymbol{N(ego(i))}|$: Anzahl der Nachbarn des Egonetzes.

Algorithmus 2 NETSIMILE's GETFEATURES

Require: $(\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\})$

1: **for all** $\boldsymbol{j} \in (\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\})$ **do**

2: $\boldsymbol{F_{G_j}} = [\,]$ // Initialisiere Merkmals-Matrix für $\boldsymbol{G_j}$

3: // berechne Merkmale für alle Knoten in $\boldsymbol{G_j}$

4: **for all** $\boldsymbol{i} \in \boldsymbol{V_j}$ **do**

5: $\boldsymbol{F_{G_j}} = \boldsymbol{F_{G_j}} \cup$

$\left\{\left\{\boldsymbol{d_i}, \boldsymbol{c_i}, \overline{\boldsymbol{d}}_{\boldsymbol{N(i)}}, \overline{\boldsymbol{c}}_{\boldsymbol{N(i)}}, \left|\boldsymbol{E_{ego(i)}}\right|, \left|\boldsymbol{E^o_{ego(i)}}\right|, \left|\boldsymbol{N(ego(i))}\right|\right\}\right\}$

6: **end for**

7: **end for**

8: // gib $\boldsymbol{node\ x\ feature}$ Matrizen zurück.

9: **return** $\{\boldsymbol{F_{G_1}}, \boldsymbol{F_{G_2}}, \dots, \boldsymbol{F_{G_k}}\}$

Tabelle 5: zeigt die genauen Formeln der einzelnen Eigenschaften, die im Algorithmus verwendet werden.

Eigenschaft	**Formel/Erläuterung**
Zahl der Nachbarknoten	k_i: Anzahl der Nachbarn eines Knoten i
Clusterkoeffizient	$C_i = \frac{2n}{k_i(k_i - 1)}$ c: Clusterkoeffizient i: untersuchter Knoten k_i: Anzahl der Nachbarn n: Anzahl der Kanten, die zwischen Nachbarn des Knoten i tatsächlich verlaufen
Durchschnittliche Zahl der Nachbarsnachbarn	-
Durchschnittlicher Clusterkoeffizient der Nachbarn	$\frac{\sum_{j=1}^{k_i} c_{ij}}{k_i}$ c_{ij}: Clusterkoeffizient des Nachbarn j des untersuchten Knoten i k_i: Anzahl Nachbarn des untersuchten Knoten i
Anzahl der Kanten des Egonetzes	$\left\|E_{ego(i)}\right\|$
Anzahl der Kanten, die aus dem Egonetz herausgehen	$\left\|E^o_{ego(i)}\right\|$
Anzahl der Nachbarn des Egonetzes	$\left\|N(ego(i))\right\|$

Nach der Berechnung der Merkmale für jeden einzelnen Knoten eines Grafen, werden diese nun aggregiert in sogenannte Signatur-Vektoren (Berlingerio et al. 2012). Dabei werden der Median, der Mittelwert, die Standardabweichung, der Symmetriekoeffizient und die Kurtosis (beschreibt, wie weit die Randbereiche einer Verteilung von einer Normalverteilung abweichen (UCLA 2022)) über alle

Werte eines Grafenmerkmals für den gesamten Graf, also über alle Knoten, gebildet und diese werden anschließend in einem Signatur-Vektor zusammengefasst (s. Tabelle 5). Jeder Graf wird nun durch sieben Signatur-Vektoren beschrieben, gemäß den sieben Eigenschaften, die für jeden einzelnen Knoten untersucht werden.

Tabelle 5: Algorithmus der Aggregator Methode. Gehe durch alle Grafen und berechne aus deren Features einen „Signatur"-Vektor pro berechnetem Feature aus dem Algorithmus in Tabelle 4. Dabei werden der Median, der Mittelwert, die Standardabweichung, die Schiefe, und die Kurtosis der Features berechnet. Die Methode gibt pro Graf einen Satz „Signatur"-Vektoren zurück.

Algorithmus 3 NETSIMILE's AGGREGATOR

Require: $\{\boldsymbol{F_{G_1}}, \boldsymbol{F_{G_2}}, \dots, \boldsymbol{F_{G_k}}\}$

1: **for all** $\boldsymbol{j} \in \{\boldsymbol{F_{G_1}}, \boldsymbol{F_{G_2}}, \dots, \boldsymbol{F_{G_k}}\}$ **do**

2: $\vec{\boldsymbol{s}}_{\boldsymbol{G_j}} = [\,]$ // Initialisiere "signature"-Vektor für $\boldsymbol{G_j}$

3: // berechne einen Satz Aggregatoren für jede Spalte in $\boldsymbol{F_{G_j}}$

4: **for all** $\boldsymbol{feat} \in \boldsymbol{F_{G_j}}$ **do**

5: $\vec{\boldsymbol{s}}_{\boldsymbol{G_j}} = \vec{\boldsymbol{s}}_{\boldsymbol{G_j}} \cup \{\boldsymbol{median}(\boldsymbol{feat}), \boldsymbol{mean}(\boldsymbol{feat}), \boldsymbol{stdev}(\boldsymbol{feat}), \boldsymbol{skewness}(\boldsymbol{feat}), \boldsymbol{kurtosis}(\boldsymbol{feat})\}$

6: **end for**

7: **end for**

8: // gib einen Satz "signature" Vektoren für den Grafen zurück.

9: **return** $\{\vec{\boldsymbol{s}}_{\boldsymbol{G_1}}, \vec{\boldsymbol{s}}_{\boldsymbol{G_2}}, \dots, \vec{\boldsymbol{s}}_{\boldsymbol{G_k}}\}$

Um die Grafen miteinander vergleichen zu können, nutzen Berlingerio et al. (2012) die Canberra-Distanz $d_{Can}(P,Q) = \sum_{i=1}^{n} \frac{|P_i - Q_i|}{P_i + Q_i}$, wobei n hier die Anzahl der zu vergleichenden Vektoren darstellt. Die Canberra-Distanz beschreibt dabei zusammenfassend in einem einzelnen Wert, wie ähnlich sich zwei Vektoren sind. In diesem Fall werden die sieben Signatur-Vektoren zweier Grafen mit deren jeweiligem Gegenstück verglichen und damit wird ermittelt, wie ähnlich sich die Strukturen der beiden zu vergleichenden Grafen sind. Durch Vergleich der beiden Signatur-Vektoren entsteht ein Wert, der die Ähnlichkeit zwischen den beiden Grafen beschreibt. Der Vergleich wird in Tabelle 6 in Form eines Algorithmus beschrieben. Der aus der Canberra-Distanz resultierende Wert wird in dieser Arbeit als mögliches Ähnlichkeitsmaß im Zusammenhang

mit dem NetSimile Algorithmus zugrunde gelegt. Wobei Werte nahe null als ähnlich betrachtet werden. Je größer die Werte sind, desto unähnlicher sind sich die Grafen. Der Quotient als solcher nimmt als höchsten Wert den Wert eins an und zwar immer dann, wenn einer der beiden Vergleichswerte null ist. Bei sieben Signaturvektoren ist damit der höchstmöglich Wert des Ähnlichkeitsmaßes sieben, da die Quotienten aufsummiert werden.

Tabelle 6: Algorithmus der Compare Methode. In dieser Methode werden die Signatur-Vektoren von verschiedenen Grafen miteinander verglichen. Die Methode gibt daraus dann einen Ähnlichkeitswert für die beiden Grafen zurück.

Algorithmus 4 NETSIMILE's COMPARE

Require: $\{\vec{s}_{G_1}, \vec{s}_{G_2}, \ldots, \vec{s}_{G_k}\}$, *doClustering*

1: **if** *doClustering* **then**
2: // gib die Cluster des angegebenen Grafen zurück.
3: **return** CLUSTER($\{\vec{s}_{G_1}, \vec{s}_{G_2}, \ldots, \vec{s}_{G_k}\}$)
4: **else**
5: // gib Ähnlichkeits-/Abstandswerte des Grafen zurück.
6: **return** PAIRWISECOMPARE($\{\vec{s}_{G_1}, \vec{s}_{G_2}, \ldots, \vec{s}_{G_k}\}$)
7: **end if**

3 Methodik

3.1 Systematische Literaturstudie zur Ermittlung vergleichbarer bereits vorhandener Arbeiten

Das dieser Arbeit übergeordnete Ziel ist die Reduzierung der Sepsis-Toten durch die Nutzung von Entscheidungsunterstützungssystemen, die auf Patientengrafen basieren. Um andere Arbeiten im Themenfeld der Darstellung von Individualpatienten in Grafen im Kontext der Grafentheorie zu identifizieren, wurde eine systematische Literaturrecherche durchgeführt (Schrodt et al. 2020). Dabei ging es darum, einen Überblick über den bisherigen Stand der Wissenschaft und über bereits etablierte Anwendungen von Grafen für individuelle Patienten zu erhalten, bei denen beispielsweise Ähnlichkeitsvergleiche durchgeführt wurden und zu ermitteln wie die temporalen Zusammenhänge in den Patientendaten für Forschungsprojekte genutzt wurden. Um Patienten vergleichen zu können, ist es in diesem Zusammenhang notwendig, dass Patientendaten durch Grafen oder Teilgrafen dargestellt werden, die den einzelnen Patienten repräsentieren. Daraus resultierten die folgenden zentralen Fragen für die Literaturrecherche:

1. Welche Arten von grafbasierten Darstellungen oder Grafdatenbanken, die Patientendaten beinhalten, sind etabliert für die Daten individueller Patienten?
2. Wie werden die Patientendaten anschließend weiterverarbeitet (beispielsweise zur Nutzung in einer Grafdatenbank oder in temporaler Modellierung oder ähnliches)?

3.1.1 Einschlusskriterien

Die untersuchten Artikel wurden auf die folgenden Einschlusskriterien hin analysiert. Das Hauptkriterium war die Nutzung des Wortes *Graf (engl. graph)* im Sinne der Grafentheorie. Die Definition von Grafen im Sinne der Grafentheorie wird vor allem charakterisiert durch das Vorhandensein von

Knoten und Kanten in einem Grafen. Diese Definition eines Grafen entspricht dem Hauptkriterium für die Nutzung des Wortes *Graf* in dieser Literaturstudie. Viele Artikel nutzten das Wort Graf in einem anderen Kontext, beispielsweise wird Graf als Synonym für Illustration oder Grafik genutzt. Solche Artikel wurden aus der weiteren Betrachtung ausgeschlossen. Es wurden außerdem Artikel ausgeschlossen, die zwar Grafen im Sinne der Grafentheorie nutzten, aber deren Grafen dann keine individuellen Patienten darstellten. Beispielsweise wurden Artikel ausgeschlossen, bei denen ein Graf einen Zusammenschluss von mehreren Patienten darstellte. Weiterhin wurden nur Artikel eingeschlossen, die auf Deutsch oder Englisch verfasst wurden. Die Datenbanksuche wurde am 20.03.2018 durchgeführt und deshalb wurden nur Artikel berücksichtigt, die bis zu diesem Datum veröffentlicht und indiziert wurden.

3.1.2 Auswahl von Artikeln

Im Rahmen dieser Literaturstudie wurde vom Verfasser dieser Arbeit die Rolle einer von vier Gutachtern übernommen. Die untersuchten Artikel wurden zur Hälfte aufgeteilt, bei jeder Hälfte wurden Titel und Abstract von je zwei Gutachtern gelesen und es wurde entschieden, ob der Artikel anhand der Einschlusskriterien eingeschlossen werden konnte oder nicht. Bei unklarem Ergebnis (ein Gutachter hat den Artikel eingeschlossen, der andere hat ihn ausgeschlossen) wurde der Artikel von einem dritten Gutachter der jeweils anderen Gruppe bewertet und daraufhin über Ein- oder Ausschluss entschieden. Anschließend wurden alle Artikel von dem Verfasser dieser Arbeit einer Volltextanalyse unterzogen und dadurch wurde erneut eine Entscheidung über Ein- und Ausschluss getroffen. Die verbleibenden eingeschlossenen Artikel wurden mit Hilfe eines MAXQDA Schemas kategorisiert, das von dem Verfasser dieser Arbeit und einem weiteren Gutachter entwickelt wurde.

Als methodische Grundlage für die systematische Literaturstudie wurden die Richtlinien von PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) (Moher et al. 2009) herangezogen.

Im Folgenden wird die Suchstrategie der Literaturstudie in wissenschaftlichen Datenbanken beschrieben. Für die Literaturstudie wurden die Stichwörter *health*

records und *graph* mit den Synonymen *medical record, patient record* und allen Pluralformen für die Datenbanksuche in den Datenbanken MEDLINE, Web of Science, IEEE Xplore and ACM digital library genutzt. Die Felder, die mit den Stichwörtern untersucht wurden waren *title* und *abstract.* Die Stichwörter wurden in die Suchanfragesyntax jeder untersuchten Datenbank übersetzt. Die spezifischen Datenbankanfragen wurden in Abbildung 3 dargestellt.

Die Artikel, die als Suchergebnisse bei der Datenbanksuche zurückgegeben wurden, wurden auf die genannten Einschlusskriterien überprüft, basierend auf den Feldern *Titel* und *Abstract.* Wenn kein Abstract vorhanden war, wurde der Volltext des Artikels genutzt. Zunächst prüften alle vier Gutachter die Einschlusskriterien an einem Beispielsatz von zehn Artikeln unabhängig voneinander. Die Ergebnisse dieses Probelaufs wurden anschließend diskutiert, um einen Konsens beim Verständnis der Einschlusskriterien zu erreichen.

Die Mitglieder jedes Teams bewerteten die Artikel, ohne dabei die Ergebnisse des anderen Teammitglieds zu kennen. Damit sollte sichergestellt werden, dass jeder Artikel zwei unabhängige Bewertungen erhält. Jeder Gutachter markierte die Artikel mit den Worten *eingeschlossen* oder *ausgeschlossen*. Für ausgeschlossene Artikel wurde der jeweilige Grund für den Ausschluss dokumentiert. Artikel, die von beiden Gutachtern eingeschlossen wurden, wurden einer Volltextanalyse unterzogen. Die Artikel, die von einem Gutachter eingeschlossen, von dem anderen aber ausgeschlossen wurden, wurden von einem dritten Gutachter bewertet, um eine finale Entscheidung für den jeweiligen Artikel treffen zu können. Der dritte Gutachter entschied damit über Ein- oder Ausschluss des betreffenden Artikels.

MEDLINE:

((("medical record"[title/abstract]) OR ("patient record"[title/abstract]) OR ("health record"[title/abstract]) OR (health record[MeSH Terms])) AND (graph[title/abstract] OR graphs[title/abstract])) OR ((("medical records"[title/abstract]) OR ("patient records"[title/abstract]) OR ("health records"[title/abstract]) OR (health records[MeSH Terms])) AND (graph[title/abstract] OR graphs[title/abstract]))

Web of Science:

(TS=((("patient record") OR ("health record") OR ("medical record")) AND (graph OR graphs))) OR (TS=((("patient records") OR ("health records") OR ("medical records")) AND (graph OR graphs)))

IEEE Xplore:

((graph OR graphs) AND ("patient record" OR "health record" OR "medical record")) OR ((graph OR graphs) AND ("patient records" OR "health records" OR "medical records"))

ACM digital library:

recordTitle:(+health +record +graph) OR recordAbstract:(+health +record +graph) OR recordTitle:(+medical +record +graph) OR recordAbstract:(+medical +record +graph) OR recordTitle:(+patient +record +graph) OR recordAbstract:(+patient +record +graph) OR recordTitle:(+health +records +graph) OR recordAbstract:(+health +records +graph) OR recordTitle:(+medical +records +graph) OR recordAbstract:(+medical +records +graph) OR recordTitle:(+patient +records +graph) OR recordAbstract:(+patient +records +graph)

Abbildung 3: Datenbanksuchanfragen für die vier verschiedenen Datenbanken MEDLINE, Web of Science, IEEE Xplore, ACM digital library. Entnommen aus: Schrodt et al. (2020)

3.1.3 Datenextraktion und Synthese

Die in den Schritten zuvor eingeschlossenen Artikel wurden anschließend einer Volltextanalyse unterzogen. Manche Artikel mussten in dieser Phase ebenfalls noch ausgeschlossen werden, da die Erfüllung der Einschlusskriterien, die in der Screeningphase noch erfüllt waren, durch die Volltextanalyse nicht bestätigt werden konnten beziehungsweise verworfen wurden. Um die Volltextanalyse zu unterstützen, wurde die computer-assisted qualitative data analysis software (CAQDAS) MAXQDA genutzt (VERBI Software GmbH 2018; Kuckartz 2014). In MAXQDA wurde eine Code-System etabliert, welches zunächst aus einem Basisartikel erzeugt wurde. In einem Code-System wurden alle zentralen Schlüsselwörter von allen untersuchten und eingeschlossenen Artikeln als hierarchische Struktur gesammelt. Jedes Schlüsselwort kann zu mehreren Artikeln zugeordnet werden und jeder Artikel kann zu mehreren Schlüsselwörtern zugeordnet werden. Das Code-System

wurde iterativ (weiter-)entwickelt, indem weitere Artikel untersucht wurden. Dafür wurden die Artikel als PDF-Dateien in die Software MAXQDA geladen, um die Informationen der Codes aus dem Code-System an die einzelnen Artikel(-abschnitte) anzuheften. Anschließend wurde das Auftreten mehrerer Codes über verschiedene Artikel hinweg untersucht und zentrale Aussagen aus dieser Analyse wurden extrahiert. Die zentralen Ergebnisse aus dieser hierarchischen Struktur sind die Unterteilung der Artikel in die Art der Grafen, die in den Artikeln genutzt wurden, die Art der Datenquellen, die Knoten- und Kanteninformationsinhalte sowie die unterschiedlichen Bearbeitungsmethoden der fertigen Grafen.

3.2 Darstellung temporaler Daten

Die Erzeugung einer Darstellungsform der temporalen Patientendaten in einem Grafen, war die zentrale Aufgabe dieser Dissertationsschrift. In den folgenden Abschnitten wird gezeigt, wie die Darstellung erarbeitet wurde und wozu sie genutzt werden soll. Dazu wird zunächst der Ausgangspunkt der Überlegungen erläutert und die in diesem Zusammenhang angefertigten Anforderungen an die Darstellung der temporalen Daten werden vorgestellt (vgl. Kapitel 3.2.1). Anschließend werden die Anforderungen der zu nutzenden Daten aus der MIMIC-III-Datenbank erfasst (vgl. Kapitel 3.2.2), um damit die für die Verarbeitung notwendigen Daten im Ergebnisteil filtern zu können (vgl. Kapitel 4.2.2). Dabei liegt der Fokus ganz klar auf der Grafdarstellung eines Patienten. Im Anschluss wird eine Vorgehensweise erläutert, die die Patientendaten aus der Grafendarstellung verarbeitet. Teil der Verarbeitung ist ebenfalls der Vergleich mehrerer Grafen und damit Patienten mit Hilfe eines Ähnlichkeitsmaßes. Das hier erarbeitete System wird abschließend in den Kontext der Entscheidungsunterstützung im klinischen Alltag eingebettet.

3.2.1 Das temporale Patientennetzwerk als Graf

Ziel der vorliegenden Arbeit ist die Schaffung einer Ausgangssituation, mit der ein Ähnlichkeitsvergleich mehrerer Patienten anhand ihrer temporalen Daten

umgesetzt in Grafen möglich ist, um möglichst ähnliche Patienten identifizieren zu können. Um dieses Ziel zu erreichen, ist es notwendig die temporalen Patientendaten in eine Form zu bringen, in der man einen möglichen Vergleich der Patienten durch ein Ähnlichkeitsmaß bewerkstelligen kann. Ausgangspunkt der Überlegung war die Darstellung der Patientendaten in Grafen. Durch die ausgeprägte Methodenvielfalt der Grafentheorie bieten sich Grafen an, um temporale Patientenprofile darzustellen und zu vergleichen. Der einfachste mögliche Graf stellt eine schlichte Abfolge von Knoten dar (vgl. Abbildung 4), eine Zeitreihe. Dabei stellt jeder Knoten einen Zeitpunkt dar, in dem teilweise mehrere Messungen am Patienten vorgenommen werden können. Diese Art der Darstellung berücksichtigt zwar die chronologische Abfolge von verschiedenen Messungen oder Symptomen, sie stellt aber weder den zeitlichen Abstand der Messungen untereinander dar noch können die Werte der Messungen oder die Kategorien einzelner Messungen untereinander verglichen werden. Die einzige Möglichkeit Patienten anhand eines solchen Zeitstrahls miteinander vergleichen zu können, wäre hier der Vergleich der Länge eines Zeitstrahls mit der Länge eines anderen Zeitstrahls, der zu einem anderen Patienten gehört.

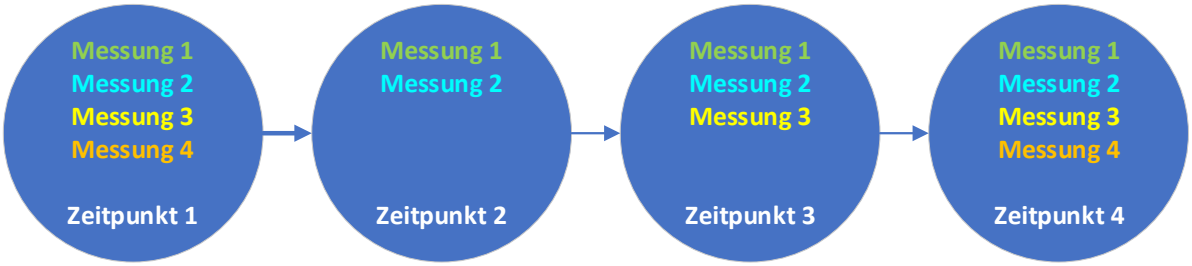

Abbildung 4: Zeitstrahl verschiedener Messungen

Als erste Anforderung an ein in dieser Arbeit zu entwickelndem temporalem Netzwerk beziehungsweise Graf kann die Notwendigkeit definiert werden, dass Messungen und andere patientenspezifische Daten in chronologischer Reihenfolge analysiert werden müssen, was durch den Grafen in Abbildung 4 bereits gegeben wäre. Das Einbinden des zeitlichen Bezugs der Daten in den Grafen kann sowohl die Diagnosen- als auch die Therapiefindung vereinfachen (Fries 1972). Als weitere Anforderung gilt allerdings, dass Messungen gleicher Kategorie untereinander vergleichbar sein sollten, was mit dem Grafen aus Abbildung 4 bereits nicht mehr umgesetzt werden kann. Das allgemeine Vorgehen bei der Untersuchung von beispielsweise Labordaten ist der Vergleich

der einzelnen Laborwerte über die Zeit (Sankaran et al. (1997)). Des Weiteren gilt es, eine Unterscheidung der Messwertkategorien in die Bewertung des Vergleichs mit einfließen zu lassen, es sollte beispielsweise die Natrium- und Kaliummessung getrennt betrachtet werden können, was mit dem Grafen aus Abbildung 4 nicht möglich ist. Diese Notwendigkeit entsteht ebenso daraus, dass Messwerte gleicher Kategorie untereinander vergleichbar sein müssen. Eine Vergleichbarkeit dieser Messwertkategorien ist nur dann möglich, wenn Messwertkategorien auch tatsächlich unterschieden werden können, beispielsweise durch deren Wert oder einer Klassifizierung des Werts. Sankaran et al. (1997) vergleichen beispielsweise direkt die Werte des Serumphosphatspiegels und dessen Entwicklung über die Zeit bei unterschiedlichen Patienten. Eine solche Bewertung wäre nicht möglich, ohne den absoluten Wertbetrag oder eine Klassifizierung des Wertes. Eine weitere Anforderung ist, dass der Graf eine Netzwerkstruktur aufweisen sollte, sodass die Grafen anhand von Werten, die ihre Topologie beschreiben, miteinander verglichen werden können. Dabei muss dieser Vergleich unabhängig von der Größe des Grafen sein, sodass auch unterschiedlich große Grafen bei ähnlicher Struktur als ähnliche Grafen erkannt werden können (Berlingerio et al. 2012). Im Zuge dieser Arbeit soll also eine grafbasierte Darstellungsform entwickelt werden, die die vorgenannten Anforderungen erfüllt (siehe auch Tabelle 7).

Tabelle 7: Anforderungen an einen Grafen basierend auf der Grafentheorie, der im Rahmen dieser Arbeit entwickelt werden soll.

Anforderungen
1. Patientenspezifische Daten sollen in chronologischer Reihenfolge dargestellt werden, um den zeitlichen Bezug der Daten mit einfließen zu lassen (Fries 1972).
2. Vergleichbarkeit von Messungen gleicher Kategorie untereinander bei verschiedenen Patienten.
3. Bewertung der Messwerte einzelner Messungen sollten in ein Ähnlichkeitsmaß mit einfließen und müssen deshalb auf dem Grafen kodiert werden.
4. Netzwerkstruktur zum Vergleich von Strukturdaten (Berlingerio et al. 2012)

Um diese Anforderungen umsetzen zu können, kann das Erzeugen der Netzwerkstruktur zunächst auf das Problem des Mapping von Zeitreihen auf Netzwerke reduziert werden (Lacasa et al. 2008; Zhang et al. 2019; Campanharo et al. 2011). Lacasa et al. (2008), Zhang et al. (2019) und Campanharo et al. (2011) gehen von Zeitreihen aus, die einen unendlich großen Satz an Messwerten beinhalten. Die Zeitreihen werden dabei fortlaufend als Funktion in einem Koordinatensystem dargestellt. Beim Mapping auf ein Netzwerk gehen dadurch Informationen verloren, da ein Netzwerk immer diskrete Punkte darstellt und es nicht möglich ist, durchgängig beschriebene Funktionen, wie im Koordinatensystem dargestellt, in einem Netzwerk darzustellen (Campanharo et al. 2011). Dadurch müssen die Kurvenfunktionen auf diskrete Punkte reduziert werden, was beispielweise durch die Teilung der Funktion in mehrere Quantile bewerkstelligt wird (Campanharo et al. 2011). Die kumulierten Werte aus diesen Quantilen resultieren anschließend in den diskreten Knoten des Netzwerks. Diese Ausgangssituation ist in der vorliegenden Arbeit allerdings nicht gegeben. Hier kann direkt von diskreten Messwerten ausgegangen werden, was das Mapping auf ein Netzwerk zunächst vereinfacht. Diese konkreten Messwerte stellen die Messwerte dar, die bei den Patientenuntersuchungen zu einem bestimmten Zeitpunkt festgeschrieben wurden. Zhang et al. (2019), Lacasa et al. (2008) und Campanharo et al. (2011) zeigen auf, dass die Knoten des resultierenden Netzwerks miteinander vernetzt werden müssen, um überhaupt eine Netzwerkstruktur und keine Aneinanderreihung von Knoten zu erhalten. In der vorliegenden Arbeit soll dabei aber nicht, wie bei Campanharo et al. (2011) gezeigt, eine recht starke Vernetzung der Knoten untereinander erzeugt werden, um damit ein Netzwerk aufzubauen, das vor allem chronologische Zusammenhänge abbildet, sondern die einzelnen Knoten sollen noch zusätzlich die Information erhalten, ob Ihre Werte im Normbereich liegen oder nicht. Durch Anforderung 3 aus Tabelle 7 musste dadurch zugunsten des Zugewinns an Informationen auf die starke Vernetzung verzichtet werden. Dieser Zugewinn wurde durch die Kategorisierung der einzelnen Messwerte und der Darstellung dieser Kategorien im Grafen erreicht. Zusätzlich trägt eine weniger starke Vernetzung der Knoten zur besseren Lesbarkeit durch einen menschlichen Leser bei (Onoue et al. 2016). Daraus resultiert ebenfalls eine Netzwerkstruktur, die mit weniger Verzweigungen auskommt und zusätzlich noch Informationen über

normale und anormale Messwerte liefern kann. Die genaue Ausarbeitung der Grafdarstellungsform ist in Kapitel 4.2.1 dargestellt. Als Vorbild dienen dabei Sichtbarkeitsgrafen (Luque et al. 2009) (engl. „visibility graphs“) vorgestellt von Lacasa et al. (2008).

3.2.2 Inhaltliche Schwerpunkte der MIMIC-III-Datenbank

Die MIMIC-III-Datenbank enthält Daten, die im Alltag einer Intensivstation anfallen. Dabei sind ganz unterschiedliche Datenarten vorhanden. So beinhaltet die Datenbank unter anderem Daten zu Vitalzeichen der Patienten, Medikationen, Labormessungen, Diagnosen, Beobachtungen, Notizen, Prozedurcodes, Berichte, die Länge des Krankenhausaufenthalts und Werte regelmäßiger Kontrollmessungen auf der Intensivstation (insbesondere die Vitalzeichen) (Johnson et al. 2016). Zunächst wurden die Datenbanktabellen herausgefiltert, die für die Erstellung eines temporalen Patientengrafen am sinnvollsten erscheinen. Dazu wurden die einzelnen Spalten auf ihre Tauglichkeit zur Nutzung in temporalen Patientengrafen untersucht anhand der in Johnson et al. (2016) beschriebenen Ausführungen. Um die jeweiligen Teile dieser Arbeit durchführen zu können, sind bestimmte Datenkategorien aus der MIMIC-III-Datenbank notwendig. So sollen alle ausgewählten Messdatenkategorien, die über den Patienten vorliegen, genutzt werden, um einen Grafen aus diesen Daten zu erzeugen. Zur Durchführung eines möglichen Vergleichs und zur Erzeugung der Grafdarstellung werden außerdem die Messwerte an sich benötigt. Damit der Ähnlichkeitsvergleich später auf ein bestimmtes Krankheitsbild hin untersucht werden kann, werden die Diagnosen benötigt. Im Anschluss an die Entwicklung der Grafdarstellung soll diese genutzt werden, um beispielhaft einen Ähnlichkeitsvergleich in Bezug auf die Sepsis durchzuführen. Außerdem sollen die Diagnosen der ähnlichsten Patienten zum aktuell ausgewählten Patienten für den Nutzer aufbereitet werden. Darüber hinaus sind Diagnosen für ein späteres Entscheidungsunterstützungssystem für die Unterstützung bei der Diagnosestellung notwendig.

Um die Daten der Grafdarstellung der einzelnen Patienten speichern zu können, muss im Anschluss an das Filtern der Datenkategorien noch eine möglichst passgenaue Datenbank ermittelt werden.

Die MIMIC-III-Datenbank besteht aus unterschiedlichen Datenbanktabellen, die in einem Krankenhausinformationssystem genutzt wurden. Hier wird beschrieben, auf welche Datenbanktabellen der Schwerpunkt in dieser Arbeit gelegt wurde und wie die Inhalte der Datenbanktabellen verknüpft sind.

Die Daten, die später für die Aufstellung eines patientenspezifischen Netzwerkes interessant sind, sind zum einen die Laborwerte, da diese gerade in Bezug auf die Sepsis entscheidende diagnostische Beiträge in Bezug auf die zugrundeliegende Infektion liefern können. Weitere wichtige Daten sind außerdem die Vitalzeichen (Blutdruck, Herzschlag et cetera). Diese Daten müssen differenziert werden nach den einzelnen Krankenhausaufenthalten. Viele Patienten haben mehr als einen Krankenhausaufenthalt in ihrer Akte verzeichnet, ausgelöst durch ganz unterschiedliche Krankheitsbilder, weswegen diese Unterscheidung notwendig sein kann. Im weiteren Verlauf der Arbeit wird allerdings ebenfalls untersucht, ob die Patientennetzwerke nach Krankenhausaufenthalten differenziert betrachtet werden müssen oder ob für einen Patienten ein Gesamtnetzwerk erstellt werden kann. Zur Kontrolle der Ergebnisse sind insbesondere die Diagnosen sowie die Notizen und Abschlussberichte von Interesse. Tabelle 8 stellt die in dieser Arbeit genutzten Tabellennamen in der MIMIC-III-Datenbank in Bezug auf ihren Inhalt dar. Die Datenbanktabellen *chartevents*, *d_items*, *d_labitems*, *icustays*, *labevents* und *patients* werden in der Ähnlichkeitsuntersuchung genutzt. Die Datenbanktabellen *d_icd_diagnoses*, *diagnoses_icd* und *noteevents* werden dagegen zur Ergebniskontrolle genutzt.

Tabelle 8: Genutzte MIMIC-III-Datenbanktabellen , deren Beschreibung und deren Zweck bei der Nutzung.

Tabellenname	Beschreibung	Zweck
chartevents	Vitalzeichen	Ähnlichkeitsuntersuchung
d_icd_diagnoses	Zuordnung eines ICD-9-Codes zu seiner Beschreibung	Ergebniskontrolle
d_items	Zuordnung der Item-ID aus Tabelle chartevents zu deren Beschreibung	Ähnlichkeitsuntersuchung
d_labitems	Zuordnung der Item-ID aus Tabelle labevents zu deren Beschreibung	Ähnlichkeitsuntersuchung
diagnoses_icd	Diagnosen eines Krankenhausaufenthalts	Ergebniskontrolle
icustays	Aufenthalte in der Notaufnahme	Ähnlichkeitsuntersuchung
labevents	Laborwerte	Ähnlichkeitsuntersuchung
noteevents	Notizen und (Abschluss-) Berichte	Ergebniskontrolle
patients	Patientenstammdaten	Ähnlichkeitsuntersuchung

3.3 Vergleich relationaler Datenbanken mit grafenbasierten Datenbanken

Bei den Herausforderungen des Social Web stoßen relationale Datenbankmodelle schnell an Ihre Grenzen, was Performance und Skalierbarkeit betrifft. Aus diesem Grund wurden sogenannte NoSQL-Datenbanken entwickelt. Unter diesen Begriff fallen unter anderem die Grafdatenbanken. Diese wurden insbesondere entwickelt um beispielsweise Beziehungen im Social Web zu erfassen, die kürzesten Wege zu finden und Besucherströme zu optimieren. Um die Grafdatenbanken untereinander zu vergleichen und eine für diese Arbeit geeignete Grafdatenbank zu identifizieren, wurden Bewertungskriterien und Ergebnisse aus der Literatur (Angles 2012) zugrunde gelegt und auf die Anwendbarkeit auf diese Arbeit untersucht. Diese Bewertungskriterien sind: die Performance der Datenbanklösungen, die Speicherart, die Nutzung einer Data Definition Language (DDL), einer Data Manipulation Language (DML) und einer Query Language, die Nutzung einer grafischen Benutzeroberfläche (GUI), die Nutzung von Knoten- und Kantenattributen in den Datenbanken und die Möglichkeit gerichtete bzw. ungerichtete Grafen zu erzeugen (Angles 2012). Die Ergebnisse der entsprechenden Untersuchung sind in Kapitel 4.3 dokumentiert.

Wurden die Patientengrafen erzeugt und gespeichert, gilt es ein passendes Ähnlichkeitsmaß für ihren Vergleich zu finden.

3.4 Auswahl eines Ähnlichkeitsmaßes

Ein Ähnlichkeitsmaß muss bestimmte Anforderungen erfüllen, um im Kontext dieser Arbeit genutzt werden zu können. Dabei gilt es zunächst, die eigentliche Funktion des Ähnlichkeitsmaßes zu erfassen. Das übergreifende Ziel dieser Arbeit, und darauf möglicherweise aufbauender Arbeiten, ist die Entwicklung eines effizienten softwaretechnischen Werkzeugs für Ärzte, mit dessen Hilfe Patienten anhand von unterschiedlichen Messwerten und Daten (Labormesswerte, regelmäßige Messwerte der Intensivstation, etc.) verglichen werden können. Daraus sollen Rückschlüsse auf eine mögliche Diagnose oder Therapie eines zu untersuchenden Patienten gezogen werden können. Der Vergleich benötigt ein Ähnlichkeitsmaß, mit dessen Hilfe die Patienten nach ihrer Ähnlichkeit zueinander sortiert werden können. So kann der Anwender möglichst ähnliche Patienten identifizieren und daraus Rückschlüsse auf Diagnose und Therapie ziehen. Damit man die Patientenprofile also vergleichen kann, benötigt man als Ähnlichkeitsmaß einen Algorithmus, der Grafen miteinander vergleicht.

Ein solches Ähnlichkeitsmaß sollte die folgenden Anforderungen erfüllen können, damit es sinnvoll eingesetzt werden kann: Zum einen sollte bei der Anwendung eines Ähnlichkeitsmaßes die Ähnlichkeitsberechnung unabhängig von der Größe der Grafen durchführbar sein, da ansonsten nur ähnlich große Grafen als strukturell ähnliche Grafen erkannt werden können. Ein Ähnlichkeitsmaß muss außerdem so ausgewählt werden, dass es auf einen Grafen angewendet werden kann, der die Anforderungen aus Kap. 3.2.1 erfüllt (vgl. Graf aus Kapitel 4.2.1). Die Ähnlichkeitsberechnung sollte außerdem auf den strukturellen Gegebenheiten eines Grafen beruhen und es sollte ebenso möglich sein, Gesamtgrafen sowie Teilgrafen mit dem Algorithmus zu untersuchen.

In Kapitel 4.4 werden für die Ausgestaltung eines Ähnlichkeitsvergleichs verschiedene Grafenvergleichsalgorithmen betrachtet und miteinander verglichen. Die Grundlage für die Auswahl des Ähnlichkeitsmaßes bildete eine unsystematische Literaturrecherche, mit der nach Algorithmen gesucht wurde, die zum Vergleich von Grafen in bisherigen Arbeiten genutzt wurden. Für alle

in Frage kommenden Kandidaten wurde im Anschluss an Ihre Auswahl eine Bewertung durchgeführt, die zu einer Entscheidung für oder gegen die Eignung der Algorithmen für die Aufgabe des Vergleichs zweier Patientengrafen führte. Ausgangspunkt der unsystematischen Literaturrecherche waren die Ergebnisse der in Kapitel 3.1 vorgestellten systematischen Literaturrecherche. Alle Artikel, die einen Grafenalgorithmus beschreiben, wurden im Rahmen der unsystematischen Literaturrecherche genutzt, um in deren Literaturverzeichnissen weitere Artikel zu finden, die grafbasierte Vergleichsalgorithmen enthielten. Dabei wurde ein rekursives Vorgehen eingesetzt und sowohl Titel als auch Abstract der Artikel wurden auf Hinweise zu einem grafbasierten Vergleichsalgorithmus durchsucht. Alle Artikel, die in Titel oder Abstract einen grafbasierten Vergleichsalgorithmus erwähnten oder andeuteten, wurden anschließend einer Volltextanalyse unterzogen. Die Algorithmen wurden anschließend in Kapitel 4.4 untereinander verglichen und auf Übereinstimmung mit den in diesem Kapitel und den in Kapitel 3.2.1 formulierten Anforderungen für Grafen hin untersucht.

3.5 Anwendung des Ähnlichkeitsmaßes

Die Patienten in der MIMIC-III-Datenbank haben beinahe alle eine oder mehrere Diagnosen für ihren jeweiligen Krankenhausaufenthalt erhalten. Demnach können Sepsis-Patienten dadurch identifiziert werden, dass ihnen die Diagnose Sepsis zugeordnet wurde. Betrachtet man nun für jeden Patienten dessen zwanzig ähnlichste Patienten, so kann man die Anzahl der Sepsis-Patienten in diesen zwanzig Patienten feststellen, indem man die Patienten mit der Diagnose Sepsis aus der Datenbank filtert.

Es wird im Folgenden angenommen, dass bei der Suche nach ähnlichen Patienten ausgehend von Sepsis-Patienten im Ergebnis ebenfalls vermehrt Sepsis-Patienten als ähnliche Patienten auftauchen sollten. Betrachtet man dagegen Nicht-Sepsis-Patienten als Ausgangspunkt, so wird angenommen, dass diese weniger Sepsis-Patienten als ähnliche Patienten nach Anwendung der Methode identifizieren sollten als das bei Sepsis-Patienten als Ausgangspunkt der Fall ist. Um diese Annahmen zu überprüfen, werden im Folgenden zwei Mittelwerte ermittelt. Der erste Mittelwert beschreibt die mittlere Anzahl an ähnlichen Sepsis-Patienten für einen Sepsis-Patienten und der andere Mittelwert beschreibt die mittlere Anzahl an ähnlichen Sepsis-Patienten für einen Nicht-Sepsis-Patienten. Die beiden Mittelwerte werden anschließend miteinander verglichen. Dadurch kann ermittelt werden, ob im Mittel durch die vorgeschlagene Methode ausgehend von Sepsis-Patienten auch wirklich mehr Sepsis-Patienten identifiziert werden können, als das zufällig durch Hintergrundrauschen der Fall ist (sprich, wenn man von Nicht-Sepsispatienten ausgeht, bei denen dann durch Zufall Sepsis-Patienten als ähnliche Patienten identifizieren).

Abbildung 5 zeigt eine schematische Darstellung der Datenerhebung für den zuvor genannten Vergleich der beiden Mittelwerte. Es wurde zunächst eine Teilmenge von 1000 Patienten zufällig aus dem MIMIC-III-Datensatz ausgewählt, die für den Test verwendet wurden. Für diese Patienten wurde nun anhand der ICD-Codes in Tabelle 16 festgestellt, ob sie die Diagnose Sepsis zugewiesen bekommen haben oder nicht. Für jeden dieser 1000 Patienten wurden nun die 20 ähnlichsten Patienten ermittelt. Zunächst wurden dabei pro

Patient alle Krankenhausaufenthalte als ein Graf zusammengefasst, später wurde derselbe Mittelwertvergleich noch einmal angewendet, dieses Mal gab es aber pro Krankenhausaufenthalt eines Patienten einen getrennten Grafen. Unter den 20 ähnlichsten Patienten wurden nun wiederum die diagnostizierten Sepsis-Fälle gezählt. So wurde in dem fiktiven Beispiel in Abbildung 5 für den Patienten 684 festgestellt, dass sich unter seinen 20 ähnlichsten Patienten 14 Sepsis-Patienten befinden, für den Patienten 1234 waren es 7. Für die beiden Nicht-Sepsis-Patienten 10 und 11 waren es dagegen vier und ein Sepsis-Patient.

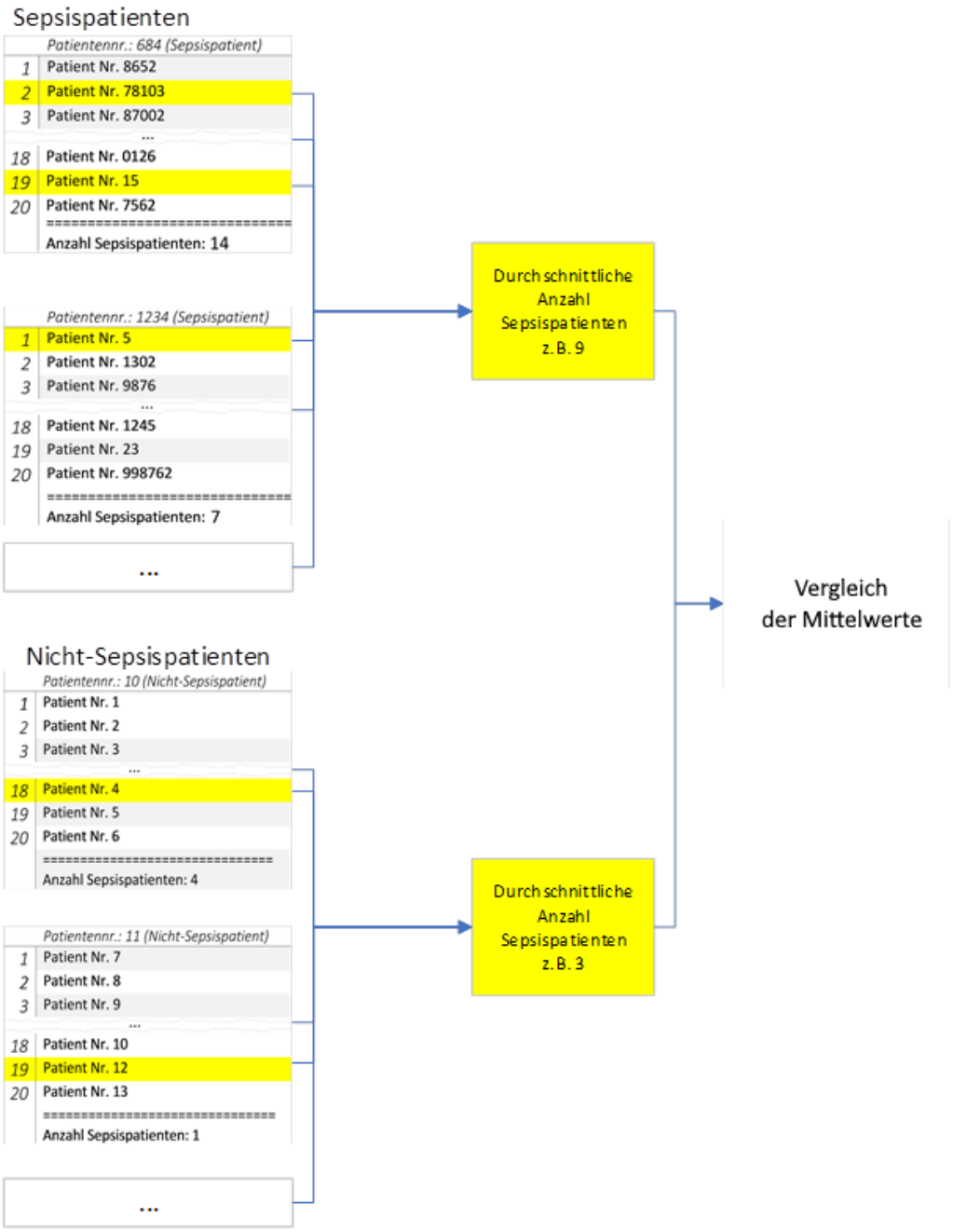

Abbildung 5: Schematische Darstellung der Datenerhebung für den Mittelwertvergleich, hier mit fiktiven Patientennummern und Ergebnissen zur Verdeutlichung der Vorgehensweise.

Für die Sepsis-Patienten und für die Nicht-Sepsis-Patienten, deren Anzahl ähnlichster Sepsis-Patienten bestimmt wurde, wird nun die Anzahl der ähnlichsten Sepsis-Patienten gemittelt. Dabei wird getrennt nach Sepsis- und Nicht-Sepsis-Patienten. Das heißt es wurde für Sepsis-Patienten die mittlere Anzahl ähnlicher Sepsis-Patient ermittelt, für Nicht-Sepsis-Patienten wurde ebenso die mittlere Anzahl ähnlicher Sepsis-Patienten ermittelt.

In Kapitel 4.5.2 wurde dann noch eine Analyse der Verteilung der Anzahl der ähnlichsten Patienten unter den zwanzig ähnlichsten Patienten vorgenommen. Dabei wurde die Zahl der Sepsis-Patienten gemessen, die unter den zwanzig ähnlichsten Patienten vorhanden waren, wenn man einen Nicht-Sepsis-Patienten und einen Sepsis-Patienten betrachtet. Die Zahl der Sepsis- und Nicht-Sepsis-Patienten wurde dabei normiert auf ihr Gesamtvorkommen in der Stichprobe, sodass ermittelt kann, zu welchem Prozentsatz mehr oder weniger ähnliche Patienten für Sepsis- oder Nicht-Sepsis-Patienten unter den zwanzig ähnlichsten Patienten waren.

4 Ergebnisse

4.1 Systematische Literaturstudie zur Ermittlung von Ansätzen zur Darstellung von Patientendaten als Graf

Im Folgenden werden die Ergebnisse der systematischen Literaturstudie vorgestellt und in Bezug zur vorliegenden Arbeit gesetzt. Eine Kurzfassung der Ergebnisse wurde publiziert (Schrodt et al. (2020)).

4.1.1 Allgemeine Ergebnisse der Literaturstudie

Die in Kapitel 3.1 beschriebene Literaturrecherche lieferte die folgenden Ergebnisse: Durch Datenbankrecherche wurden in MEDLINE 201 Treffer gefunden, 107 in Web of Science, 58 in IEEE Xplore und 92 in ACM digital library. Nach der Eliminierung der Dubletten verblieben noch insgesamt 383 Artikel, die den Anforderungen der Datenbankeingaben gerecht wurden. Nach Anwendung der Einschlusskriterien waren sich die Gutachter bei 320 der 383 einige über Ein- oder Ausschluss der Artikel (84%). Für 63 Abstracts gab es gegenläufige Meinungen, weshalb für diese Artikel die Meinung eines dritten Gutachters notwendig und ausschlaggebend war. Letztendlich wurden 42 Abstracts durch übereinstimmende Meinungen der ersten beiden Gutachter eingeschlossen, sechs weitere Abstracts wurden durch einen Drittgutachter eingeschlossen, der die Konfliktartikel in den Abstracts und in den Volltexten (falls notwendig) analysierte und dann für oder gegen den Artikel entschied. Insgesamt wurden so 48 Artikel eingeschlossen (12,5 %), 335 (87,5 %) wurden ausgeschlossen. Die Hauptgründe für den Ausschluss waren, dass

- die Artikel das Wort *Graf* nicht im Sinne der Grafentheorie nutzten und
- Artikel, die Grafen im Sinne der Grafentheorie nutzten, die nicht auf individuelle Patienten abgebildet wurden.

Viele der Artikel wurden aus der weiteren Betrachtung ausgeschlossen, weil sie von vorneherein schon den Begriff des Grafen in einem anderen Kontext verwendet hatten. Bei den meisten Artikeln, die so ausgeschlossen wurden,

wurde ein Bild oder ein Diagramm als Graf bezeichnet, statt eines Grafen mit Knoten und Kanten im Sinne der Grafentheorie. Zum anderen bezogen sich die Anforderungen darauf, dass der Graf einen individuellen Patienten oder eine Teilmenge von Daten eines individuellen Patienten darstellen sollte und dass die Artikel in Deutsch oder Englisch verfasst wurden (Schrodt et al. 2020).

Nachdem diese 48 eingeschlossenen Artikel in der der Volltextanalyse untersucht wurden, verblieben noch elf Artikel (2,9 % der Gesamtmenge der gefundenen Artikel) (Atif et al. 2007; Liu et al. 2015; Campbell et al. 2015; Chen et al. 2016; Esteban et al. 2015; Kaur und Rani 2015; Müller et al. 1996; Puentes et al. 2008; Zhang et al. 2017; Zhang et al. 2016; Hanzlicek et al. 2004) für die Code-Struktur-Analyse in MAXQDA.

Dabei konnten bei sechs der elf Artikel, die den Anforderungen entsprachen, Grafen identifiziert werden, die einen temporalen Zusammenhang zwischen Daten eines individuellen Patienten herzustellen versuchten. Die Fokussierung temporaler Zusammenhänge in den untersuchten Artikeln, konnte auch durch das gleichzeitige Auftreten von temporalen Beziehungen innerhalb der Grafen mit Diagnose und Labordaten als Knoten in den untersuchten Grafen gezeigt werden.

4.1.2 Code-Schema

In Kapitel 3.1 wurde bereits die Entwicklung des Code-Schemas beschrieben, welches genutzt wurde, um die Artikel zu kategorisieren. Dieses Code-Schema wurde in MAXQDA umgesetzt und in Abbildung 6 dargestellt. Die Hauptkategorien nach Analyse der verbleibenden elf Artikel waren *a) Datenquelle (engl. data source) b) Gesamtzweck (engl. overall purpose / function) c) Graf Eigenschaften (engl. graph properties) d) untersuchte Krankheiten (engl. investigated disease) e) technische Weiterverarbeitung der Grafen (engl. technical processing of graph).* Zunächst wurden die Methoden der Darstellung und der Speicherung der Daten in Grafen und Grafdatenbanken untersucht. Diese Untersuchung kann den Kategorien *data source* und *graph properties* zugeordnet werden. Diese Codierungen definieren wie Grafen generiert werden und was deren Inhalte in Knoten und Kanten sind. Das zweite Hauptuntersuchungsfeld betraf die weitere Verarbeitung der Grafen und deren Inhalte. Demnach konnten die Verarbeitungsmethoden dieser Grafen und deren Inhalte durch die Codierungen *overall purpose / function* und vor allem *technical processing of graph* abgebildet werden. Dieser zweite Schritt half dabei zu verstehen wie bereits existierende Studien Patientengrafen verarbeiten und welche Ziele diese Untersuchungen mit der Verarbeitung der Patientengrafen verfolgten.

- Codesystem [211]
  - categorization [0]
  - heterogeneous data mining [1]
  - causal networking [2]
  - database/data structure approach, SNOMED [1]
  - structure representation [2]
  - temporal event data mining [6]
- data source [0]
  - free-text records [1]
  - paper health records [1]
  - image-based information [2]
  - UNMC de-identified clinical research database [1]
  - SNOMED CT clinical findings [2]
  - Electronic Health Record (EHR) [11]
  - healthcare information system [2]
- overall purpose / function [0]
  - drug investigation [1]
  - information gaining [3]
  - quality improvement [2]
  - population management [1]
  - data mining [1]
  - data warehouse [1]
  - predictive modeling [2]
  - disease diagnosis [2]
  - disease pattern [1]
  - patient segmentation [2]
  - personalized medicine [6]
- graph [1]
  - patient graph [4]
  - node content [0]
    - chemical structure of drug [1]
    - words [1]
    - observation [1]
    - emotion [1]
    - procedures [2]
    - functional vertices [4]
    - matter [1]
    - anatomic parts [3]
    - patient [2]
    - diagnosis [7]
    - Electronic Health Records [1]
    - clinical notes [1]
    - problems [2]
    - medications [5]
    - vital signs [1]
    - laboratory data [7]
  - edge content [0]
    - causal relation [2]
    - properties [1]
    - word link [1]
    - patient drug prior associations [1]
    - drug similarity [1]
    - patient similarity [1]
    - number of combinations between nodes [0]
    - Anatomo-Functional relations [1]
    - matter relation [1]
    - spatial relations [2]
    - taxonomical relation [1]
    - status and date [1]
    - temporal relationships [6]
  - graph properties [0]
    - undirected [0]
    - directed [8]
    - unweighted graph [0]
    - weighted graph [5]
      - average duration [2]
    - content of subgraphs [0]
      - detected phenotypes [1]
  - supervision of information [0]
    - supervised information [2]
    - semi-supervised information [2]
    - unsupervised information [1]
- investigated diseases [0]
  - (childhood) leukemia [1]
  - coronary heart diseases (CHD) [1]
  - COPD [1]
  - hypercholesterolemia [1]
  - diabetes [1]
  - kardiovascular disease [1]
  - kidney failure [1]
  - geriatric diseases [1]
  - brain tumors [1]
  - pleural effusion [1]
  - Congestive Heart Failure (CHF) [2]
  - CHF with COPD precondition [2]
  - pneumonia due to an influenza [1]
  - parainfluenza virus [1]
  - breast cancer [2]
- investigated diseases ICD [0]
  - C71.* | D33.* | D43.* brain tumors [1]
  - C50.* breast cancer [2]
  - C95.* (childhood) leukemia [1]
  - I25.* coronary heart diseases (CHD) [1]
  - I25.* | I51.9 kardiovascular disease [1]
  - I50.9 Congestive Heart Failure (CHF) [2]
  - I50.9 | J44.* CHF with COPD precondition [2]
  - J44.* COPD [1]
  - J90.* pleural effusion [1]
  - J10.0 pneumonia due to an influenza [1]
  - J12.2 parainfluenza virus [1]
  - E78.0 hypercholesterolemia [1]
  - E11.* diabetes [1]
  - N19.* kidney failure [1]
  - geriatric diseases [1]
- investigated diseases zusammengefasst [0]
  - cancer [4]
  - heart disease [6]
  - lung disease [6]
  - E78.0 hypercholesterolemia [1]
  - E11.* diabetes [1]
  - N19.* kidney failure [1]
  - geriatric diseases [1]
- processing of graph (technical) [0]
  - predictive modeling [1]
  - kind of storage [0]
    - GraphML [2]
    - graph database [1]
      - MongoDB [1]
      - neo4j [2]
    - SNOMED [2]
    - RDF [1]
  - conversion [0]
    - SNOMED to neo4j [1]
  - clustering [2]
  - similarity comparison of graphs [3]
- Sets [0]

Abbildung 6: Code-System erzeugt in MAXQDA , entnommen aus Schrodt et al. (2020)

Grafeigenschaften

In Tabelle 9 werden alle Arten von Knoteninhalten der elf eingeschlossenen Artikel dargestellt. Sechs der Artikel nutzten Labordaten, die in den Knoten der Grafen dargestellt wurden, fünf der Artikel nutzten hier Medikationen und Diagnosen. Funktionelle Knoten wurden vier Mal verwendet, während anatomische Knoten und Patientenprobleme jeweils zwei Mal zum Einsatz kamen. Prozeduren, Vitalzeichen und Patientenknoten wurden jeweils einmal genutzt. Patientenknoten sind Knoten, die dazu dienen einen Patienten in einem Grafen zu identifizieren.

Tabelle 9: Überblick über alle verschiedenen Knoteninhalte in den untersuchten Artikeln. Spalte 2 zeigt die Anzahl der Artikel, in denen der Knoteninhalte aus Spalte 1 genutzt wurde; Entnommen aus Schrodt et al. (2020)

Node content	**# papers**
laboratory data	**6**
medications	**5**
diagnoses	**5**
functional nodes	**4**
anatomic nodes	**2**
patient problems	**2**
procedures	**1**
vital signs	**1**
patient nodes	**1**

Im Gegensatz dazu zeigt Tabelle 10 den Inhalt der Kanten, wie sie in den elf untersuchten Artikeln eingesetzt wurden. In zwei Artikeln repräsentierten die Kanten kausale Beziehungen, sodass die Knoten, die durch die Kanten verbunden wurden, in kausalem Zusammenhang miteinander standen. In einem Artikel zeigten die Kanten anatomisch-funktionale Zusammenhänge, wohingegen in zwei Artikeln räumliche Beziehungen durch die Kanten dargestellt wurden. Im Detail zeigten die Kanten räumliche Verbindungen zwischen Gehirnarealen. *Taxonomische Beziehungen* (engl. taxonomical relations) und *Status und Datum* (engl. status and date) sind zwei weitere Inhaltsarten, die für die Kanten genutzt wurden. Beide wurden jeweils in einem der untersuchten Artikel genutzt. Die Kanteninhalte, die in den meisten Artikeln genutzt wurde, sind *temporale Beziehungen* (engl. temporal relations). Diese

wurden in sechs der untersuchten Artikel eingesetzt. Dabei war die Interpretation der temporalen Beziehung, die durch die Kanten dargestellt wurden in den Artikeln etwas unterschiedlich, allerdings stellten diese so dargestellten temporalen Beziehungen immer einen zeitlichen Zusammenhang zwischen den verbundenen Knoten dar.

Tabelle 10: Überblick über alle verschiedenen Kanteninhalte in den Artikeln, entnommen aus Schrodt et al. (2020)

Edge content	**# papers**
Causal relations	**2**
Anatomic-functional relations	**1**
Spatial relations	**2**
Taxonomical relation	**1**
Status and date	**1**
Temporal relations	**6**

4.1.3 Graftypen

Tabelle 11 zeigt alle Typen von Grafen, die in den Artikeln genutzt wurden um elektronische Patientendatensätze (engl. Electronic Medical Records (EMR)) eines einzelnen Patienten darzustellen. Die meisten der verbleibenden Artikel nutzten die Darstellung eines EMR in einem Graf zur Darstellung eines einzelnen Patienten in einem temporalen Zusammenhang (Liu et al. 2015; Chen et al. 2016; Esteban et al. 2015; Kaur und Rani 2015; Müller et al. 1996; Zhang et al. 2016; Zhang et al. 2017; Zhang et al. 2016). Im Gegensatz dazu wurden kausale Netzwerkzusammenhänge in zwei Artikeln genutzt um Patientendaten darzustellen (Kaur und Rani 2015; Müller et al. 1996). Heterogenes Data-Mining wurde durch einen Artikel genutzt und beschreibt die Darstellung von sehr verschiedenen Datenarten eines Patienten in einem einzelnen Grafen (Müller et al. 1996) während Datenbank- beziehungsweise Datenstrukturansätze in zwei Artikeln genutzt wurden. Diese Artikel demonstrierten mögliche Methoden, Patientendaten in einer Grafdatenbank oder einer grafähnlichen Struktur zu speichern (Campbell et al. 2015; Hanzlicek et al. 2004). Zwei weitere Artikel nutzten Grafen für die strukturelle Repräsentation von Gewebearealen im Gehirn (Atif et al. 2007; Puentes et al. 2008).

Tabelle 11: Überblick über Grafkategorien , die in den elf Artikeln genutzt wurden, entnommen aus Schrodt et al. (2020).

Graph category	Frequency	Source
Causal networking	2	(Kaur und Rani 2015; Müller et al. 1996)
Heterogeneous data mining	1	(Müller et al. 1996)
Database / data structural approach	2	(Campbell et al. 2015; Hanzlicek et al. 2004)
Structure representation	2	(Atif et al.; Puentes et al. 2008)
Temporal event data mining	6	(Chen et al. 2016; Zhang et al.; Zhang et al.; Liu et al. 2015; Esteban et al. 2015; Müller et al. 1996)

4.1.4 Datenquellen

Des Weiteren wurden verschiedene Datenquellen untersucht, die in den verschiedenen Artikeln für Patientendaten genutzt wurden (s. Tabelle 12). EMRs stellen den größten Teil der Datenquellen, die in den Artikeln genutzt wurden (62,5 %). Manche Artikel nutzten außerdem Informationen aus Bildgebungsverfahren (12,5 %) oder Daten von Informationssystemen für das Gesundheitswesen (engl. health care information system) (12,5 %). Ein Artikel untersuchte Daten aus *SNOMED CT clinical findings* und ein weiterer eine wissenschaftliche Datenbank (jeweils 6,25 %).

Tabelle 12: verschiedene Datenquellen für Patientendaten , die in den unterschiedlichen Artikel genutzt wurden.

Data sources	Frequency	Percentage	References
Image-based information	2	12,5	(Atif et al.; Puentes et al. 2008)
University of Nebraska Medical Center (UNMC) de-identified clinical research database	1	6,25	(Campbell et al. 2015)
SNOMED CT clinical findings	1	6,25	(Campbell et al. 2015)
Electronic Health Record	10	62,5	(Liu et al. 2015; Campbell et al. 2015; Chen et al. 2016; Esteban et al. 2015; Zhang et al.; Zhang et al.; Zhang et al.; Puentes et al. 2008; Müller et al. 1996; Kaur und Rani 2015; Hanzlicek et al. 2004; Zhang et al. 2017; Zhang et al. 2016; Zhang et al. 2017)
Healthcare information system	2	12,5	(Liu et al. 2015; Puentes et al. 2008)

4.1.5 Verarbeitung der Grafen

Tabelle 13: Überblick über alle genutzten Arten der Verarbeitung von Patientengrafen in den eingeschlossenen Artikeln.

Kind of processing	# papers
Prognostic modelling	**1**
Storing of graphs	**5**
Similarity comparison of graphs	**2**
Presentation of patient data	**9**

Tabelle 13 zeigt die Anzahl der Artikel, die die aufgelisteten Arten der Verarbeitung der resultierenden Grafen mit Patientendaten nutzten. Nur in einem Artikel wurde demnach der Grafen zu Prognosezwecken verwendet (Kaur und Rani 2015). Fünf Artikel untersuchten das Speichern von Patientengrafen in

verschiedenen Arten und Weisen, beispielsweise in Grafdatenbanken, während die Autoren in zwei weiteren Artikeln daran interessiert waren, Ähnlichkeitsvergleiche der erzeugten Grafen anzustellen. Dagegen spielt in neun Artikeln die Darstellung der Patientendaten eine zentrale Rolle. In diesen Artikeln wurde dann von einer weiteren Verarbeitung abgesehen.

4.1.6 Ziele und Inhalt der untersuchten Artikel

Die Forschungsziele, die in den verschiedenen Artikeln beschrieben wurden, wichen im Detail sehr voneinander ab, aber ein Anwendungsfall, der oft in den Artikeln angesprochen wurde, war die personalisierte Medizin. Diese Begrifflichkeit trat in sechs der elf untersuchten Artikel auf. Weitere Kernziele der Artikel waren Qualitätsverbesserungen, Informationsgewinnung, prädiktive Modellierung (engl. predictive modelling), Diagnose von Krankheiten und Patienten Segmentierung (alles wurde jeweils zwei Mal in den Artikeln benutzt). Außerdem wurden noch die Ziele des Populationsmanagement, Data-Mining, Data Warehouse und die Erkennung von Krankheitsbildern angesprochen, welche jeweils in einem der elf Artikel genutzt wurden.

Um diese Zielsetzungen zu erreichen, verfolgten die Autoren der elf Artikel sehr unterschiedliche Strategien. Atif et al. (2007) nutzten beispielsweise bildbasierte Informationen von Gehirnen, um eine grafbasierte zerebrale Beschreibung der Gehirnanatomie zu erzeugen. Dieser räumliche Graf wurde händisch für jeden einzelnen Patienten erzeugt und anschließend konnten die Patienten durch Nutzung der jeweiligen Grafen miteinander verglichen werden (Atif et al. 2007). Im Gegensatz dazu nutzten Campbell et al. (2015) das SNOMET CT Konzeptmodell in einer Grafdatenbank-Architektur, um sich die Ontologie und Polyhierarchie von SNOMED CT zunutze zu machen, welche es schwierig macht EMR-Daten in relationalen Datenbanken darzustellen (Campbell et al. 2015). Daher wurde der Versuch unternommen, statt der relationalen Datenbanken eine Grafdatenbank zu verwenden. In den so erzeugten Grafen werden SNOMED CT Daten in einem spezifisch erzeugten Grafformat gespeichert. Im Gegensatz dazu ist die Risikoprävention das Hauptziel von Chen et al. (2016), weshalb die Autoren einen grafbasierten Lernalgorithmus

entwickelten, um dieses Ziel zu erreichen (Chen et al. 2016). Um die klinische Entwicklung eines individuellen Patienten mit Nierenversagen zu modellieren, entwickelten Esteban et al. (2015) die Basis für ein zukünftiges klinisches Entscheidungsunterstützungssystem. Dieses grafbasierte Modell enthält tausende von klinischen Ereignissen wie Labordaten, verordnete Tests und Diagnosen (Esteban et al. 2015) und stellt einen Patienten in einem Grafen dar. Von der Grundidee kommt dieser Artikel nahe an die Idee der vorliegenden Arbeit heran, jedoch ist die Herangehensweise zu differenzieren. Hanzlicek et al. (2004) beschrieben MUDR EHR, einen multimedial verteilten Gesundheitsdatensatz für die Entscheidungsunterstützung. Dieser EMR enthält mehrere medizinische Konzepte, die dabei helfen sollten, einen Patienten in einer strukturierten Art und Weise beschreiben zu können, anders als bei den bisherigen Freitext-Datensätzen. Die Darstellung der Patientendaten als strukturierte Datensätze erleichtert die Verarbeitung dieser Datensätze durch automatische Systeme (Hanzlicek et al. 2004). Kaur und Rani (2015) beschreiben dagegen ein Modell, das verschiedene Datenspeicher von Patientendaten miteinander kombiniert. In dieser Architektur erzeugt der Anwender seine Anfrage als Query vergleichbar mit einer SQL-Anfrage in einer grafischen Benutzeroberfläche und die Architektur unter dieser Oberfläche übersetzt die Anfrage in eine Anfrage des betreffenden Datenspeichers für diese Anfrage. Dabei kommen ganz unterschiedliche Datenspeicher zum Einsatz, unter anderem auch relationale Datenbanken. Die Grundidee dabei war, ein einheitliches System zu schaffen, in dem der Anwender in einer standardisierten Anfragesprache Daten laden kann, ohne auf die unterschiedlichen Anfragesprachen der einzelne untergeordneten Datenspeicher zurückgreifen zu müssen (Kaur und Rani 2015). Der resultierende Graf in diesem Artikel hilft dabei die richtigen Informationen aus diesen Datenspeichern für den Anwender zu extrahieren. Liu et al. (2015) nutzten dagegen longitudinale Patientendaten, um einen sogenannten temporalen Graf zu erzeugen. Diese Grafen wurden in verschiedene Phänotypencluster überführt, sodass diese Phänotypen die Leistung der Diagnostizierung verbessern sollten (Liu et al. 2015). Der resultierende Graf repräsentiert einen Patienten und seine klinischen Ereignisse in einem zeitlichen beziehungsweise temporalen Kontext. Der Ausgangspunkt von Müller et al. (1996) war das Fehlen eines klinischen Kontextes in anderen

Ansätzen. Die Autoren lösten das Problem, indem Sie einen graf-grammatikalischen (engl. graf-grammar) Ansatz entwickelten, um ein graforientiertes Patientenmodell zu erzeugen, das es erlaubt, den klinischen Kontext darzustellen (Müller et al. 1996). Puentes et al. (2008) nutzten ähnlich wie Atif et al. (2007) Grafen, um Informationen von bildbasierten Informationen des Gehirns zu gewinnen, um die räumlichen Beziehungen von hirn-anatomischen Eigenheiten von individuellen Patienten zu modellieren. Dieser Ansatz wird speziell für die räumliche Modellierung von zerebralen Tumoren genutzt (Puentes et al. 2008). Dagegen entwickelten Zhang et al. (2017) ein *faltendes neuronales Netzwerk* (engl. convolutional neural network) auf Basis verschiedener Attribute des Patienten (beispielsweise Diagnosen, Verfahren und Medikationen). Dabei nutzten Sie einen Grafen, der seine Daten von EMR-Daten bezieht (Zhang et al. 2017). Zhang et al. (2016) erzeugten dagegen eine vereinheitliche Darstellung von EMRs eines individuellen Patienten in einem zeitlichen Zusammenhang. Im zweiten Schritt wurde durch Nutzung des Grafs mit einem modifizierten Algorithmus ein temporales Profil jedes Patienten erstellt. Dieser Ansatz wurde dann zur Risikoprävention genutzt (Zhang et al. 2016).

4.2 Darstellung temporaler Daten bzw. konzeptionelle Darstellung der Patientendaten als Graf

Die grundsätzlichen Anforderungen an die hier zu entwickelnde Grafenstruktur aus medizinischen Patientendatensätzen wurden in Kapitel 3.2.1 bereits festgestellt. So soll ein Graf die chronologische Abfolge der Messwerte codieren, außerdem sollen verschiedene Messgrößen unterschieden werden können (beispielsweise Chlorid und Natriummessung sollen in Laborwerten unterschieden werden können). Des Weiteren muss der Graf eine Netzwerkstruktur aufweisen. Durch die reine chronologische Abfolge entsteht eine lineare Zeitreihe, die, wie bereits in Kapitel 3.2.1 festgestellt, lediglich über deren Länge mit anderen Zeitreihen verglichen werden kann. Ein ausschließlicher Längenvergleich der einzelnen Teilgrafen ist im

Zusammenhang mit dieser Arbeit aber nicht zielführend. Das Grundproblem bei rein chronologischen Abfolgen beziehungsweise Zeitstrahlen wird in Abbildung 7 dargestellt. Hier wurde ein und derselbe Graf auf drei unterschiedliche bildliche Arten und Weisen dargestellt. Während in Abbildung 7 A alle fünf Knoten in einer linearen Ebene sitzen, sind die Knoten in Abbildung 7 B versetzt und in Abbildung 7 C ineinander verdrillt. Trotzdem steckt aus grafentheoretischer Sicht unter der Voraussetzung, dass die Kanten in den einzelnen Abschnitten dieselbe Länge haben, genau derselbe Informationsgehalt in allen drei Grafen. Um aus diesen Zeitstrahlen Informationen bezüglich der Vergleichbarkeit der Grafen zu gewinnen, müssen die Informationen der Messwertgrößen abstrahiert werden und in die Grafdarstellung mit einfließen. Das kann nur erreicht werden, wenn die Zeitstrahlen zu Netzwerken transformiert werden. Lacasa et al. (2008) gehen wie oben erläutert in ihren Sichtbarkeitsgrafen den Weg, dass sie Knoten zusätzlich miteinander verbinden, indem sie die Messwerte als Balken darstellen, die ihren Ausschlag in y-Richtung in einem kartesischen Koordinatensystem aus der absoluten Messwertgröße beziehen. Balken im Sichtfeld eines betrachteten Balkens werden dann mit dem betrachteten Balken verbunden. In der vorliegenden Arbeit wird dagegen eine andere Art der Abstraktion der Zeitstrahlen in Netzwerke entwickelt. Hier wird durch zusätzliche Kategorisierung der Messwerte in normale und anormale Messwerte und die zusätzliche Einführung von chronologischen Abfolgen innerhalb dieser Kategorien eine Netzwerkstruktur erzeugt.

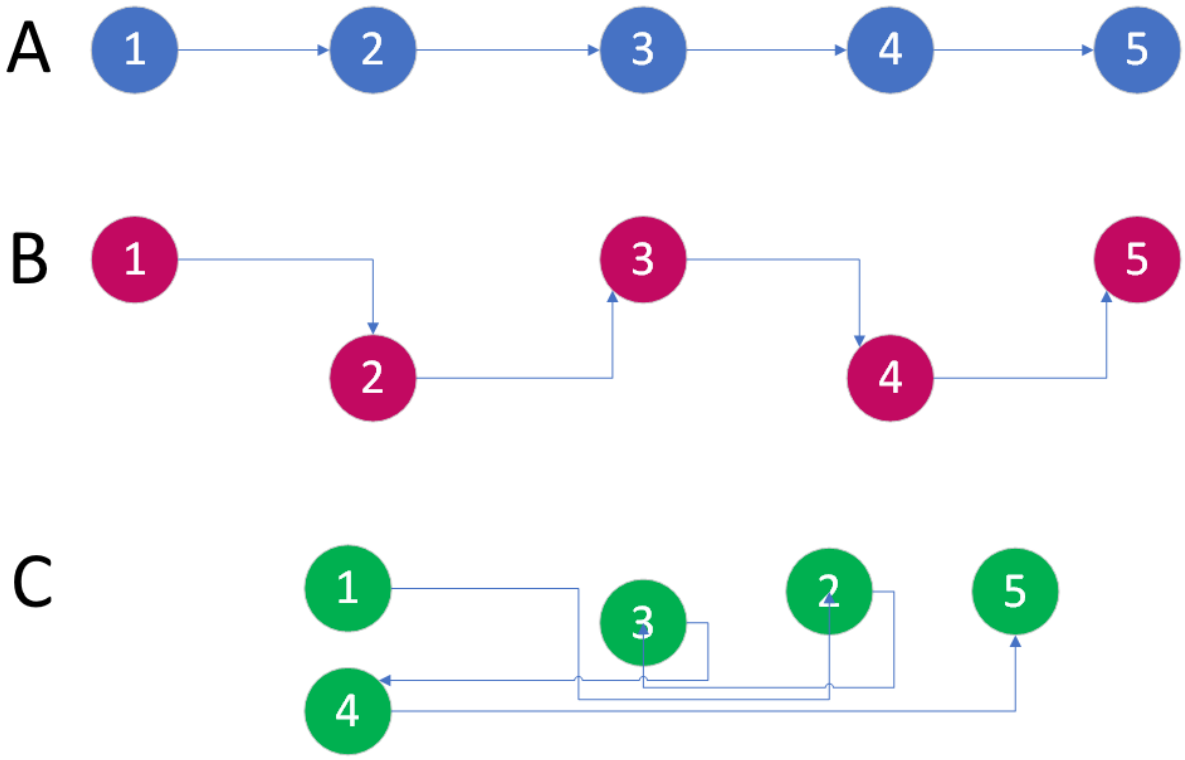

Abbildung 7: Drei Darstellungen desselben Grafen. Graf A, B und C zeigen alle drei ein und denselben Graf, der lediglich bildlich anders dargestellt wurde.

Um diese Anforderungen umsetzen zu können, kann das Erzeugen der Netzwerkstruktur zunächst auf das Problem des Mapping von Zeitreihen auf Netzwerke reduziert werden (Lacasa et al. 2008; Zhang et al. 2019; Campanharo et al. 2011). Lacasa et al. (2008), Zhang et al. (2019) und Campanharo et al. (2011) gehen von Zeitreihen aus, die einen unendlich großen Satz an Messwerten beinhalten. Die Zeitreihen werden dabei fortlaufend als Funktion in einem Koordinatensystem dargestellt. Beim Mapping auf ein Netzwerk gehen dadurch Informationen verloren, da ein Netzwerk immer diskrete Punkte darstellt und es nicht möglich ist, durchgängig beschriebene Funktionen in einem Netzwerk darzustellen (Campanharo et al. 2011). Dadurch müssen die Kurvenfunktionen auf diskrete Punkte reduziert werden, was beispielweise durch die Teilung der Funktion in mehrere Quantile bewerkstelligt wird (Campanharo et al. 2011). Die kumulierten Werte aus diesen Quantilen resultieren anschließend in den diskreten Knoten des Netzwerks. Diese Ausgangssituation ist in der vorliegenden Arbeit allerdings nicht gegeben. Hier kann direkt von diskreten Messwerten ausgegangen werden, was das Mapping auf ein Netzwerk zunächst vereinfacht. Zhang et al. (2019), Lacasa et al. (2008) und Campanharo et al. (2011) zeigen auf, dass die Knoten des resultierenden

Netzwerks miteinander vernetzt werden müssen, um überhaupt eine Netzwerkstruktur und keine Aneinanderreihung von Knoten zu erhalten. In der vorliegenden Arbeit soll dabei aber nicht, wie bei Campanharo et al. (2011) gezeigt, eine recht starke Vernetzung der Knoten untereinander erzeugt werden, um damit ein Netzwerk aufzubauen, das vor allem chronologische Zusammenhänge abbildet, sondern die einzelnen Knoten sollen noch die Information erhalten, ob Ihre Werte im Normbereich liegen oder nicht. Daraus resultiert ebenfalls eine netzwerkartige Struktur, die mit weniger Verzweigungen auskommt und zusätzlich noch Informationen über normale und anormale Messwerte liefern kann. Diese Grafdarstellung wurde durch die Sichtbarkeitsgrafen aus Luque et al. (2009) inspiriert.

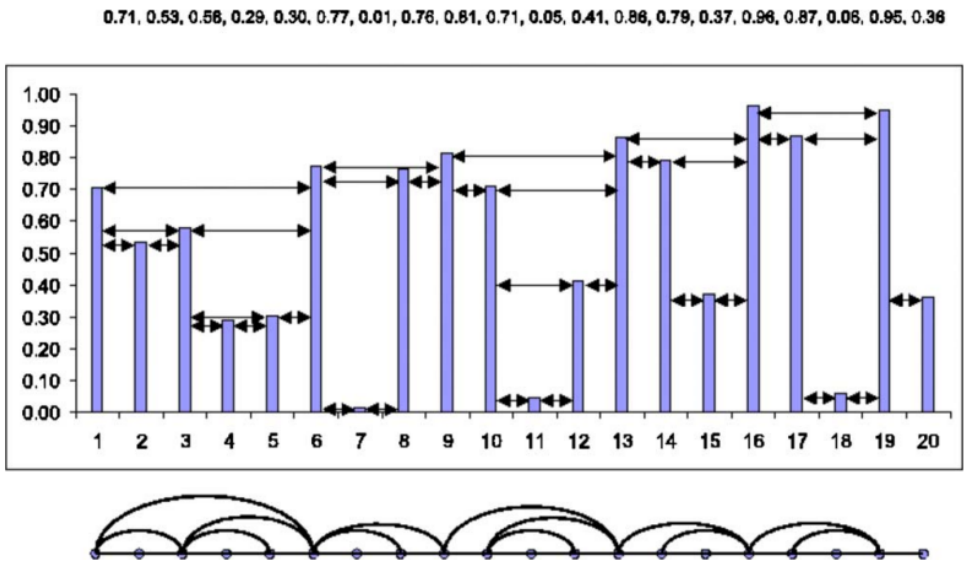

Abbildung 8: Der Sichtbarkeitsgraf. Illustratives Beispiel eines horizontalen Sichtbarkeitsalgorithmus. Im oberen Bereich wurde eine Zeitreihe dargestellt, während im unteren Bereich ein Graf dargestellt wurde, der durch den horizontalen Sichtbarkeitsalgorithmus generiert wurde. Jedes Datum in der Zeitreihe entspricht einem Knoten im Grafen, zwei Knoten werden miteinander verbunden, wenn deren korrespondierende Datenhöhe (Höhe der Balken in diesem Beispiel) größer sind als alle Datenpunkte zwischen ihnen (Luque et al. 2009). Entnommen aus Luque et al. (2009), Fig. 2.

Abbildung 8 zeigt beispielhaft wie der Sichtbarkeitsalgorithmus funktioniert. Im oberen Teil ist dabei eine Zeitreihe dargestellt, im unteren der daraus berechnete Graf, der mit Hilfe eines Sichtbarkeitsalgorithmus ermittelt wurde. Jeder Datenpunkt stellt einen Knoten im Grafen dar und zwei Datenpunkte werden dann miteinander verbunden, wenn deren Messwerte größer sind als die korrespondierenden Messwerte zwischen diesen beiden Werten. Die Grundidee

von Lacasa et al. (2008) und deren Sichtbarkeitsgrafen ist letztendlich die Überführung von Zeitstrahlen beziehungsweise zeitlichen zusammenhängenden Daten in Grafen. Dieser Grundidee soll in dieser Arbeit gefolgt werden. Die Umwandlung von strukturierten Patientendaten in einen Grafen zur Anwendung von Vergleichsalgorithmen der Grafentheorie ermöglicht allerdings eine etwas andere Herangehensweise. Der in Abbildung 9 dargestellte und in dieser Arbeit entwickelte Graf unterteilt die Messwerte zunächst in zwei Gruppen: normale und anormale Messwerte. Die Daten der MIMIC-III-Datenbank enthalten bereits teilweise die Information über normale und anormale Messwerte, insbesondere in den Labordaten. Durch die gewählte Speichermethode in neo4j als Grafdatenbank (Kapitel 4.3.2) oder auch in der PostgreSQL-Datenbank (Kapitel 4.3.1 und Kapitel 4.3.3) werden allerdings auch – zusätzlich zu der Darstellung in Abbildung 9 – Informationen zu den absoluten Messwerten sowie zum zeitlichen Abstand der Knoten gespeichert (s. Abbildung 10). Wie in Kapitel 4.5 dargestellt, reicht für bestimmte Anwendungen bereits die kategorisierte Form der Messwerte aus, um passable Ergebnisse zu erhalten. Je nach Anwendungsfall kann durch die flexible Speicherung der detaillierteren Daten dann aber auch auf die absoluten Werte zurückgegriffen werden. Diese Werte bestimmen dann zum einen die Länge der Kanten, zum anderen aber auch den Ausschlag einzelner Knoten in einer gedachten Nulllinie für Normwerte (s. Abbildung 10). Die eingeführte Kategorisierung der Messwerte in normale und anormale Messwerte bietet an dieser Stelle die Möglichkeit, eine Netzwerkstruktur zu bilden. Anders als bei den Sichtbarkeitsgrafen (Lacasa et al. 2008) wird hier nicht die absolute Größe der Messwerte und deren Sichtfeld zum nächsten nicht verdeckten Messwert als Grundlage der Bildung einer Netzwerkstruktur zugrunde gelegt, sondern die Möglichkeit, durch die Kategorisierung der Messwerte eine zusätzliche chronologische Abfolge einzuführen. Demnach werden, zusätzlich zur Kategorie unabhängigen chronologischen Reihenfolge, noch zwei weitere chronologische Abfolgen eingeführt, die jeweils innerhalb normaler Messwerte beziehungsweise der anormalen Messwerte platziert werden. Betrachtet man nun dieselbe Messgröße (beispielsweise „Chlorid“) bei unterschiedlichen Patienten, so können normale und anormale Messwerte über die Zeit schwanken. So können beispielsweise bei einem Patienten alle Messwerte im Normbereich liegen, bei einem zweiten

Patienten könnten dagegen alle Messwerte als anormal betrachtet werden. Bei einem dritten Patienten könnten zunächst normale Messwerte in der zeitlichen Abfolge dargestellt werden, dann anormale, beispielsweise verursacht durch medikamentenbedingte Nebenwirkungen. Bei einem vierten Patienten könnten die Messwerte ein periodisches Auf und Ab anzeigen. Durch die unterschiedlichen Verzweigungsmuster von normalen und anormalen Messwerten untereinander, können Unterschiede in den einzelnen Messgrößen festgestellt werden und somit ist ein Vergleich zwischen verschiedenen Patientengrafen möglich.

In Abbildung 9 wird ein solcher Graf dargestellt. Dieser Graf erfüllt zusätzlich alle Anforderungen, die in Kapitel 3.2.1 an einen entsprechenden Grafen gestellt wurden und kann daher als Ausgangspunkt für das weitere Vorgehen betrachtet werden. Die tatsächlichen zeitlichen Abstände zwischen zwei Knoten, genauso wie die absoluten Messwertgrößen, können mit den vorgestellten Speichermethoden im Grafen verankert werden, so können diese Daten bei Bedarf zusätzlich zum Vergleich der Patientengrafen herangezogen werden. Welche der Möglichkeiten dabei genutzt wird, entscheidet sich dabei vor allem durch den gewählten Vergleichsalgorithmus. Ein entsprechender Graf wird in Abbildung 10 dargestellt.

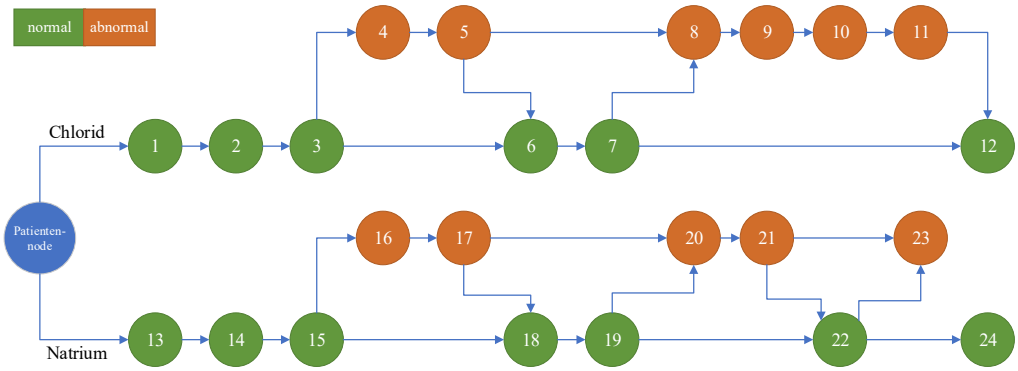

Abbildung 9: Beispielnetzwerke eines Patientenprofils für den Chlorid-Wert und den Natriumwert einer Blutuntersuchung. Im oberen Zweig wird der Chlorid-Wert dargestellt, im unteren der Natriumwert. Für jeden Zweig werden die Knoten in chronologischer Reihenfolge dargestellt. Die grünen Knoten sind dabei Werte im Normbereich, die orangenen sind Werte im anormalen Bereich. Die Knoten im normalen und im anormalen Bereich werden jeweils chronologisch verbunden und die Knoten über beide Bereiche hinweg werden auch noch einmal chronologisch miteinander verbunden. Alle Zweige sind letztendlich durch den Patientenknoten verbunden. Jeder Patient erhält so ein eigenes Netzwerk.

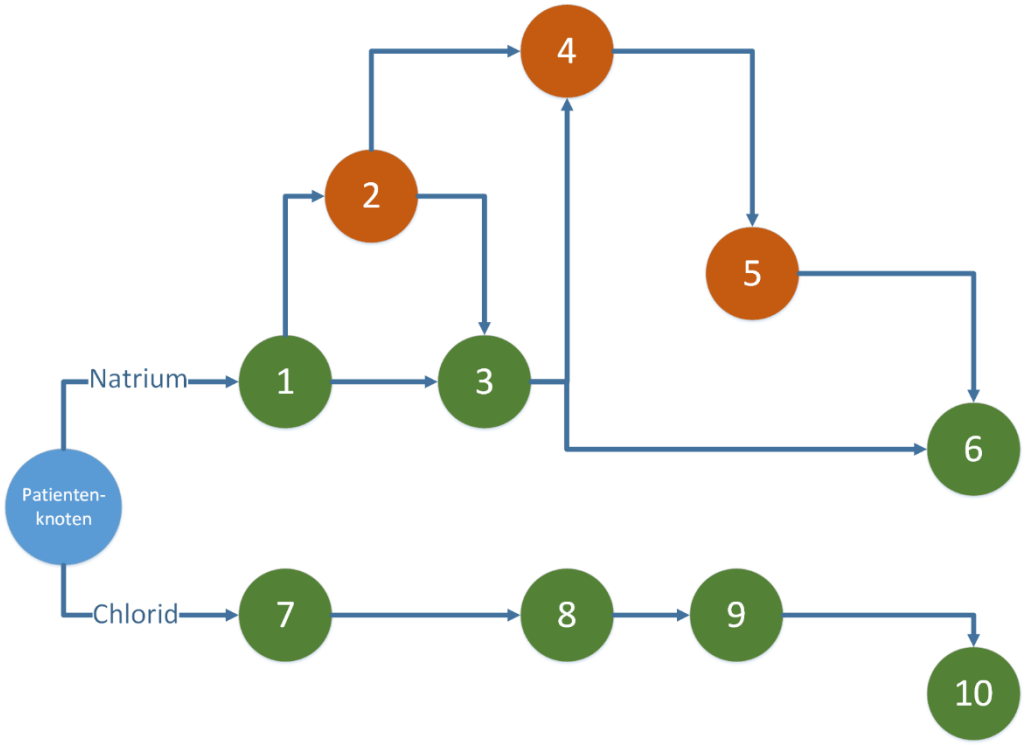

Abbildung 10: Graf mit Verwendung der originalen Messwerte : Beispielnetzwerk eines Patientenprofils für den Chlorid- und Natriumwert einer Blutuntersuchung. Für jeden Zweig werden die Knoten in chronologischer Reihenfolge dargestellt. Die orangefarbenen Knoten stellen auch hier die anormalen Werte dar, die grünen Knoten stellen die normalen Werte dar. Durch Speicherung der Messwerte und zeitlichen Abstände kann ein Graf wie hier dargestellt werden. Durch die Verwendung der originalen Messwerte entsteht eine weitere Vergleichsebene, die durch die y-Position der Messwerte gebildet wird.

Abbildung 9 zeigt ein Netzwerk, das den Anforderungen aus Kapitel 3.2.1 gerecht wird. Die temporalen Daten werden chronologisch dargestellt. Ausgehend von einem Patientenknoten (blau) wird in Abbildung 9 pro Kategorie von Messungen ein Zweig des Grafen pro Patient erstellt. Dabei sind in diesem Beispiel zwei Kategorien von Messungen genannt, nämlich die Chlorid-Messung sowie die Natrium-Messung im Blut. Jede dieser beiden Kategorien erhält einen eigenen Zweig für den gezeigten Patienten. In dem Zweig wird pro Messung ein Knoten dargestellt. Die Knoten werden hintereinander dargestellt und bilden damit die chronologische Reihenfolge der Messungen ab. Grüne Knoten stellen dabei Messwerte im Normbereich dar, orangefarbenen Knoten stellen Werte außerhalb des Normbereichs dar. Die Anforderung nach dem Einfluss der Messwerte auf die Ähnlichkeitsanalyse ist damit gegeben. Damit eine auswertbare Netzwerkstruktur aus diesen Daten entstehen kann, wurden die Knoten der normalen Werte pro Zweig sowie die Knoten der anormalen Werte pro Zweig jeweils chronologisch verbunden. Außerdem werden die Knoten des anormalen Bereichs mit den Knoten des normalen Bereichs noch einmal chronologisch verbunden. Auf diese Weise erhält man ein Teilnetzwerk für jede Messkategorie, welche über den Patientenknoten miteinander verbunden werden. Da auf der Intensivstation oft in regelmäßigen Intervallen die gleiche Messung durchgeführt wird, ist der zeitliche Abstand einzelner Messungen hier vernachlässigbar. Damit ist rein die chronologische Reihenfolge als temporaler Aspekt dieses Netzwerks ausschlaggebend. Durch die gezeigte Struktur des Netzwerks lassen sich allerdings über Netzwerkstrukturanalysen Aussagen über gleichbleibende oder sich stetig ändernde Werte machen. Im Gesamten betrachtet, erhält man so ein Profil eines Patienten basierend auf seinen Messwerten, welches durch Kombination der einzelnen Messkategorien und deren Struktur die Möglichkeit bietet, um mit anderen Patientennetzwerken gleicher Art verglichen werden zu können.

Abbildung 10 zeigt einen ähnlich gearteten Grafen, der sich aber die zeitlichen Abstände der Knoten zunutze macht. Aus diesem zeitlichen Abstand wird die Kantenlänge zwischen zwei Knoten abstrahiert. Es wird ebenfalls abstrahiert dargestellt, wie unterschiedliche absolute Messwerte in den Knoten dargestellt werden können. Diese beiden Zusatzinformationen können in einem dafür

ausgelegten Vergleichsalgorithmus ebenfalls genutzt werden, um eventuell noch bessere Ergebnisse zu erzielen.

Gerade für die Laborwerte aus der MIMIC-III-Datenbank bietet sich ein Graf wie in Abbildung 9 oder Abbildung 10 an. Die Messwerte der Laborwerte werden in der MIMIC-III-Datenbank direkt mit einer Kennzeichnung versehen, ob der Messwert im Normbereich liegt oder außerhalb des Normbereichs. Dadurch lässt sich für die Messwerte ohne spezielle Konvertierungen ein Graf nach diesem Schema erzeugen.

Die Darstellung von Patientendaten als Netzwerk in einem Grafen stellt letztendlich das Ergebnis für Ziel 1 dar. Ziel 1 fokussiert sich auf die Erarbeitung einer möglichen grafenbasierten Darstellungsform für temporale Ereignisse im Krankheitsverlauf eines Patienten, sprich Messwerte und ähnliche Ereignisse, deren temporaler Zusammenhang durch die Grafdarstellung erhalten bleiben soll. Diese Darstellung als Graf dient der formalen Repräsentation von klinischen Daten in einer Fallbasis für klinische Entscheidungsunterstützung.

4.3 Effiziente Speicherung von Grafen

Die in Kapitel 4.2.1 entwickelte konzeptionelle Darstellungsform eines Patienten als Graf, soll nun in effizienter Art und Weise digital gespeichert werden. Dazu werden in diesem Abschnitt zunächst relationale mit grafenbasierten Datenbanken verglichen und anschließend grafenbasierte Datenbanken untereinander verglichen. Daraus wird abgeleitet, welche Datenbankart sinnvoll für die Fragestellung dieser Arbeit genutzt werden soll.

4.3.1 Vergleich relationaler und grafenbasierter Datenbanken

Die Nutzung von Grafdatenbanken hat bestimmte Vorteile gegenüber relationalen Datenbanken. So können Grafdatenbanken sehr viel mehr Daten schneller verarbeiten als relationale Datenbanken. Das liegt daran, dass Grafdatenbanken durch Nutzung einer entsprechenden Abfragesprache den Fokus auf den einzelnen Knoten setzen. Relationale Datenbanken dagegen untersuchen bei einer Abfrage die gesamte Datenbank (Jaiswal und Agrawal 2015). Gerade wenn viele Join-Befehle in den relationalen Datenbanken genutzt werden müssen, um verschiedene Datenbanktabellen zu verbinden, kommt die relationale Datenbank an ihre Performancegrenze (Vicknair et al. 2010).

Die Nutzung relationaler Datenbanken bietet im ersten Moment also eindeutige Nachteile gegenüber der Nutzung grafenbasierter Datenbanken, weshalb im folgenden Abschnitt unterschiedliche Grafdatenbanken auf Ihre Eignung untersucht werden.

Bei genauerer Betrachtung der Grafdarstellung der vorliegenden Arbeit konnte allerdings festgestellt werden, dass sich die vermeintlichen Nachteile zumindest bei der hier vorgestellten Grafstruktur so nicht bestätigen lassen. In der Literatur wird bei der Beschreibung der Vorteile von Grafdatenbanken gegenüber relationalen Datenbanken oft von stark vernetzten Grafen ausgegangen (Miller 2013). Grafen dieser Art zeigen zwischen den Knoten der einzelnen Grafen viele Verbindungen auf, wobei in der vorliegenden Arbeit pro Knoten maximal zwei eingehende und zwei ausgehende Verbindungen möglich sind. Die

hochvernetzten Netzwerke beziehen sich dabei allerdings oft auf klassische Beispiele wie den sozialen Graf (soziale Netzwerke), Empfehlungssysteme (Vorschläge verwandter Artikel beispielsweise beim Onlineshopping) oder biologische Netzwerke (beziehungsweise Netzwerke der Bioinformatik) (Miller 2013), wobei hier stets von einigen Hundert Kanten pro Knoten ausgegangen werden kann. Eine mögliche andere Erklärung könnte die Art der Nutzung der Daten darstellen. Die meisten Anwendungen von hochvernetzten Grafen, die sich auf die dargestellten Vorteile der Grafdatenbanken beziehen, betrachten das Finden eines einzelnen Grafen und dessen Verzweigungen. Bei den meisten grafentheoretischen Vergleichsalgorithmen bezieht sich der Algorithmus (hier beispielsweise NetSimile) allerdings auf die Topologie des Gesamtgraphen (Berlingerio et al. 2012). Diese Art von Algorithmus muss demnach jeden einzelnen Knoten untersuchen und die zugehörigen Statistiken berechnen. Eine derartige Vorgehensweise könnte demnach in einer Matrixberechnung zum Beispiel in relationalen Datenbanken effizienter umgesetzt werden, als es durch Nutzung eines Grafdatenbankmodells möglich ist, da hier der Fokus bei den meisten Anwendungen vor allem auf dem Finden bestimmter Knoten und deren Verzweigungen liegt. Zusätzlich wird in der vorliegenden Arbeit von einem Grafen pro Patienten ausgegangen, wobei zwischen den einzelnen Patientengrafen keine Verbindung vorhanden ist. Im Gegensatz dazu wird beispielsweise bei den sozialen Grafen meist nur ein einziger vernetzter Graf betrachtet (Ugander et al. 2011). Da ein möglicher Vergleichsalgorithmus aber trotzdem grafentheoretische Methoden anwenden könnte, um die Grafen miteinander zu vergleichen, wird im Folgenden neben der relationalen Datenbank auch eine Grafdatenbank untersucht. Dazu wird im folgenden Schritt ein geeignetes Grafdatenbankmodell gesucht. Das ausgewählte Grafdatenbankmodell dient dann als Grundlage für den Vergleich mit der relationalen Datenbank, im Speziellen mit dem Datenbank-Management-System (DBMS) PostgreSQL. Die PostgreSQL-Datenbank wurde deshalb als relationale Datenbank gewählt, weil sie recht ressourcenschonend arbeitet (Riggs 2015), weil PostgreSQL ein OpenSource-Datenbankmanagementsystem ist (Eisentraut und Helmle 2013) und weil die MIMIC-III-Datenbank unter anderem in einem Paket von CSV-Dateien ausgeliefert wird mit beigelegtem Skript, mit dem man die Daten in eine PostgreSQL-Datenbank einlesen kann (Johnson et al. 2016).

Bei der ausgewählten Grafdatenbank wurde sich anschließend auch die strukturelle Anforderung an CSV-Dateien für einen Bulkimport zunutze gemacht, um die PostgreSQL-Datenbank strukturell ähnlich einer Grafdatenbank aufzubauen. Dazu wurden die Kanten zwischen den relational dargestellten Knoten durch Speicherung der Kanten in weiteren Datenbanktabellen nachgeahmt. Diese so konvertierte Datenbank wurde in Kapitel 4.4 testweise dazu verwendet, einen beispielhaft angewandten Vergleichsalgorithmus zu nutzen.

4.3.2 Vergleich grafenbasierter Datenbankmodelle

Die bekanntesten Vertreter der Grafdatenbanken sind Neo4j, HypergraphDB, DEX, InfoGrid, Sones und VertexDB (Angles 2012). Vergleicht man die Grafdatenbanken Neo4j, HypergraphDB und DEX im Hinblick auf ihre Performance, so konnte gezeigt werden, dass Neo4j und DEX die effizientesten dieser drei Datenbanken sind (Dominguez-Sal et al. 2010). Angles (2012) stellt neun verschiedene Grafdatenbanken gegenüber. Dabei werden unterschiedliche Speicherarten betrachtet: Hauptspeicher, externer Speicher, Backendspeicher und zusätzlich wird noch die Möglichkeit der Nutzung von Indizes betrachtet. Dabei decken die meisten Datenbanken die Speicherarten Hauptspeicher, Externer Speicher und auch die Indizes ab, während nur Filament, HyperGraphDB und vertexDB Backendspeicher ermöglichen. Backendspeicher ist für die Anwendung in dieser Arbeit allerdings nicht notwendig, weshalb nur Hauptspeicher, externer Speicher und die Indizes betrachtet werden. Nur Neo4j, HyperGraphDB, DEX und AllegroGraph nutzen die Indizes und diese beiden Speicherarten. HyperGraphDB hat allerdings, wie oben bereits beschrieben, eine schlechtere Performance als Neo4j und DEX, weshalb es hier aus der weiteren Betrachtung ausgeschlossen wird. Laut Angles (2012) haben Neo4j und DEX zum Stand von 2012 weder eine Data Definition Language (DDL), eine Data Manipulation Language (DML) noch eine Query Language, AllegroGraph besitzt diese drei Sprachanteile allerdings. Außerdem kann bei allen drei Datenbanken eine Programmierschnittstelle (API) verwendet werden, eine Grafische Benutzeroberfläche (GUI) hat allerdings nur AllegroGraph zum Stand

von 2012 (Angles 2012). Heute nutzt Neo4j die an SQL angelehnte Sprache Cypher (Johnpaul und Mathew) als DML und Query Language. Neo4j nutzt außerdem seit geraumer Zeit den Neo4j Browser als GUI für seine Datenbank. DEX heißt heute Sparksee und nutzt auch heute noch weder eine DDL, DML, eine Query Language noch eine GUI (Angles et al. 2013), weshalb Sparksee aus der weiteren Betrachtung ausgeschlossen wird. Angles (2012) stellt ebenfalls die Art der Grafen gegenüber. So werden die Grafen der Datenbank AllegroGraph als einfache Grafen bezeichnet, bei denen die Nodes zwar beschriftet sind, es aber keine Node-Attribute gibt. In AllegroGraph gibt es außerdem Grafen mit gerichteten Kanten und beschrifteten Kanten, aber keine Kantenattribute. Neo4j wird dagegen als Attributgraf bezeichnet, wobei hier die Knoten beschriftet sind und Knotenattribute angelegt werden können. Die Kanten sind gerichtet, beschriftet und es gibt die Möglichkeit Kantenattribute zu vergeben (Angles 2012). Damit hat die Grafdatenbank Neo4j eindeutige Vorteile gegenüber AllegroGraph und wird in dieser Arbeit als Grafdatenbank genutzt. Genau diese Speicherung der Attribute ermöglicht außerdem die Nutzung der absoluten Messwerte der Knoten und die Speicherung der Zeitabstände zwischen den Knoten wie sie beide in Kapitel 4.2.1 beschrieben und in Abbildung 10 dargestellt werden.

4.3.3 Einlesen der MIMIC-III-Daten in die Grafdatenbank

Im Zuge dieser Arbeit werden Daten aus der MIMIC-III-Datenbank ausgelesen und in die Grafdatenbank neo4j eingelesen. Dies erfolgt mit Hilfe einer eigens dafür entwickelten C# WPF-Anwendung. Dazu wird zunächst pro Patienten, der eingelesen werden soll, ein Knoten in der Grafdatenbank erzeugt, der die für die folgenden Untersuchungen wichtigsten Eigenschaften des Patienten enthält. Diese Eigenschaften sind das Geburtsdatum, das Todesdatum, ein Zeitstempel und eine ID. Das Geburtsdatum des Patienten ist notwendig, um dessen Alter feststellen zu können. Dadurch können beispielsweise Säuglinge von Erwachsenen unterschieden werden. Das Todesdatum ist wichtig, um zu ermitteln, ob und in welchem Krankheitszusammenhang der Patient verstorben

ist. Der Zeitstempel ist für die weitere Bearbeitung des Patienten notwendig. Sollten nämlich neue Messwerte für den Patienten hinzukommen, muss er neu eingelesen werden. Durch den Zeitstempel können solche intensiven Rechenvorgänge vermieden werden, sollten keine Änderungen gemacht worden sein. Die ID dient der eindeutigen Identifikation des Knoten. Außerdem werden für alle Knoten noch die Minuten seit dem ersten Patientenereignis mit berechnet. Für den Patientenknoten wird hier ein Wert von -1 hinterlegt. Dadurch kann der Patientenknoten von den restlichen Knoten unterschieden werden.

Nach Erzeugen des Patientenknoten werden die Daten der Ereignisse des Patienten in die Grafdatenbank eingelesen. Zunächst werden nur die Labor-Daten (Tabelle Labevents in der MIMIC-III-Datenbank) in neo4j übernommen, im Anschluss daran dann die Vitalwerte des Patienten (Tabelle Chartevents in der MIMIC-III-Datenbank). Die Laborwerte konnten recht einfach eingelesen werden, da für sie bereits in der MIMIC-III-Datenbank eine Kennzeichnung vergeben wurde, ob der Messwert des Laborwertes im normalen Bereich liegt oder nicht. Für die Vitalwerte war diese Zuordnung nicht ganz so einfach, da für die Vitalwerte eine Kennzeichnung der Messwerte wie sie in den Laborwerten vorhanden ist, nicht gespeichert wurde. Diese Schwierigkeit wird im Abschnitt „Vitalwerte (Chartevents)“ in diesem Kapitel dargestellt.

Um die Daten aus technischer Sicht in die Neo4j Datenbank einlesen zu können wurde der von Neo4j bereitgestellte Neo4j-Treiber (Neo4j Driver) und der Neo4jClient genutzt. Diese konnten direkt über C# angesprochen werden und durch Nutzung der Abfragesprache Cypher konnte so auf die Daten in der Grafdatenbank zugegriffen werden.

Um die Daten aus der PostgreSQL-Datenbank auslesen zu können, wurde Npgsql genutzt, das direkt in ein C#-Projekt als DLL integriert werden kann und so den Zugriff auf die PostgreSQL Datenbank erlaubt.

Labor-Daten (Labevents)

In den Labordaten wird der Bezug zum jeweiligen Patienten über die Spalte subject_id ermittelt. In dieser Spalte steht die Patientennummer. Für den ersten Eintrag für den Patienten wird der Zeitstempel festgehalten. Dieser Zeitpunkt wird als Zeitpunkt null für den Patienten betrachtet. Alle weiteren Ereignisse stehen in Bezug zu diesem Zeitpunkt und werden in Minuten seit Zeitpunkt null angegeben. Jedes Ereignis in den Laborwerten bekommt seinen eigenen Knoten. So wird beispielsweise für jede Chloridmessung aus dem Blutbild ein Knoten erzeugt. Dieser Knoten bekommt dann also einen Wert, der die Minuten seit Zeitpunkt null darstellt. An den Knoten werden außerdem noch die Werte der Spalten „itemid“, „charttime“, „value“, „valueuom“, und „valuenum“ übergeben. Des Weiteren erhält der Patientenknoten noch die Patientennummer und die Knotenart „labevent“. Die Spalte „itemid“ aus der MIMIC-III-Datenbank enthält eine eindeutige Nummer für die Art des Laborwertes. So hat Chlorid beispielsweise die itemid 50806. Durch „value“ und „valueuom“ erhält die Grafdatenbank den Messwert und die dazugehörige Einheit. „valuenum“ enthält dagegen den Messwert als Zahl und „flag“ enthält bei den Laborwerten die Information, ob der Wert im Normbereich oder außerhalb des Normbereichs lag. Die Information über den Normbereich wird nicht als extra Eigenschaft an den Knoten übergeben, sondern als Teil des Knotennamens mit in den Knoten eingefügt. Der Knotenname wird für Werte im Normbereich wie folgt zusammengesetzt: Event_<item_beschriftung>. Für Werte außerhalb des Normbereichs wird dieser wie folgt zusammengesetzt: Event_<item_beschriftung>_abnormal. Die „item_beschriftung“ wird für die itemid aus der Tabelle d_labitems aus der Spalte „label“ geladen.

Im Anschluss an die Erstellung der Patienten- und der Ereignisknoten müssen diese miteinander verbunden werden, um einen temporalen Zusammenhang der einzelnen Knoten herzustellen. Dabei wird der temporale Zusammenhang auf zwei unterschiedliche Arten und Weisen hergestellt. Zunächst werden die ersten Knoten jeder Messwertart (eine Messwertart wird definiert über dieselbe itemid) mit dem Patientenknoten verbunden. Die ersten Knoten einer Messwertart haben immer die kleinste Zahl von Minuten zum ersten Messzeitpunkt und können dadurch identifiziert werden. Die Beschriftung der so geschaffenen Kanten ist

„START“. Im nächsten Schritt werden alle Knoten einer Messwertart chronologisch miteinander verbunden. Diese Kanten bekommen die Beschriftung „CHRONOLOGICALLY_FOLLOWING“. Dass die Methoden der Grafentheorie anwendbar sind, muss nun aus diesen von einem Zentrum wegführenden Linien ein Netzwerk entstehen. Dieses Netzwerk wird dadurch erzeugt, dass für dieselbe Messwertart Knoten im Normbereich untereinander noch einmal chronologisch verbunden werden und Knoten außerhalb des Normbereichs ebenfalls nochmal miteinander verbunden werden. Die so entstehenden Kanten werden mit „SAME_GROUP_CHRONOLOGICAL“ beschriftet.

Vitalwerte (Chartevents)

Die Vitalwerte bzw. Chartevents für den einzelnen Patienten werden aus der Datenbanktabelle chartevents geladen und anhand der Spalte subject_id identifiziert. In dieser Spalte steht die Patientennummer. Um den Zeitpunkt des ersten Eintrags des Patienten festzuhalten, wurde zunächst der erste Eintrag identifiziert, der für den Patienten existiert. Dieser wurde als Zeitpunkt null festgehalten. Der Zeitpunkt null ist zunächst nicht unbedingt derselbe wie der Zeitpunkt null der Laborwerte. Später wurden die Zeitpunkte dann zusammengefasst, sodass alle Zeitpunkte, die für den Patienten berechnet wurden auf diesen Nullpunkt zurückgehen. Es wird also für alle Werte der Abstand zum Nullpunkt berechnet und im Feld „min_from_first_charttime“ festgehalten. Jedes Ereignis in den Vitalwerten bzw. Chartevents bekommt seinen eigenen Knoten. So wird beispielsweise für jede Messung des Blutdrucks eines Patienten ein eigener Knoten erzeugt. An diesen Knoten werden noch die Werte der Spalten „itemid“, „charttime“, „value“, „valueuom“ und „valuenum“ übergeben. Jeder Knoten erhält außerdem ein Feld „Art“, in das bei den Chartevents der Wert „chartevent“ eingetragen wurde, während bei den Laborwerten der Wert „labevent“ eingetragen wurde. Dadurch können Knoten von Laborwerten von Knoten für Chartevents unterschieden werden. Die Spalte „itemid“ aus der MIMIC-III-Datenbank enthält in der Tabelle d_items eine eindeutige Nummer für die Art des Chartevents. So bezeichnet die Nummer 211

beispielsweise die Herzfrequenz. Durch „value“ und „valueuom“ erhält die Grafdatenbank den Messwert und die dazugehörige Einheit. „valuenum“ enthält dagegen den Messwert als Zahl. Die Patienten- und die Ereignisknoten wurden getrennt voneinander erstellt und müssen nun, wie auch die Ereignisknoten untereinander, miteinander verbunden werden. Das ist notwendig, um einen temporalen Zusammenhang zwischen den Knoten herstellen zu können. Zunächst werden die ersten Knoten jeder Messwertart (eine Messwertart definiert sich über dieselbe itemid) mit dem Patientenknoten verbunden. Die so geschaffene Kante erhält die Beschriftung „START“. Der Wert min_from_first_charttime legt fest, wie viele Minuten vom ersten Zeitpunkt einer Messung vergangen sind. Derjenige Knoten, der pro Messwertart den kleinsten Wert hat, ist der Startknoten, der mit dem Patientenknoten verbunden wird. Nun werden alle Knoten einer Messwertart chronologisch untereinander verbunden, angefangen beim Startknoten pro Messwertart. Die so entstehenden Kanten erhalten die Beschriftung „CHRONOLOGICALLY_FOLLOWING“. Um die Methoden der Grafentheorie anwenden zu können, muss aus dieser Ansammlung von Knoten nun eine Netzwerkstruktur entstehen. Dazu werden nun zusätzlich zur chronologischen Reihenfolge in der gesamten Messwertart weitere Kanten hinzugefügt. Und zwar werden Werte im Normbereich und Werte außerhalb des Normbereichs derselben Messwertart getrennt voneinander betrachtet. Die chronologische Reihenfolge der Messwerte im Normbereich sowie der außerhalb des Normbereichs wird getrennt voneinander festgelegt. Die so entstehenden Kanten erhalten die Beschriftung „SAME_GROUP_CHRONOLOGICAL“. Da in den Chartevents deutlich mehr aufgezeichnete Ereignisse pro Patient gespeichert wurden als in den Laborevents, wurde bei den Chartevents das Einlesen der Ereignisknoten und das Verbinden der Knoten in unterschiedlichen Funktionen getrennt voneinander programmiert. Dadurch können zunächst alle Knoten erzeugt werden und im Anschluss daran chronologisch verbunden werden. Die große Menge der Knoten machte es außerdem notwendig, die Erzeugung der Knoten und Kanten im Multithreadingbetrieb der Anwendung laufen zu lassen. Bei Nutzung des Multithreading werden die Rechenlasten auf die unterschiedlichen Kerne der Rechnerhardware und deren Threads verteilt. Dadurch können mehrere Berechnungen gleichzeitig ausgeführt werden, was die Rechendauer um beinahe

diesen Faktor reduzieren sollte. Abbildung 11 stellt den Ablauf und die Nutzung des Multithreading beim Erzeugen der Kanten in den Chartevents schematisch dar. Der Anwender kann zunächst einen Bereich von Patientennummern vorgeben, die eingelesen werden sollen. Für jeden Patienten wird nun ein Thread erstellt, bis die Maximalanzahl an Threads erreicht wurde. Dadurch können die Daten der Patienten parallel berechnet werden, gegenüber der sequentiellen Datenverarbeitung kann so enorm Zeit eingespart werden. In jedem Thread werden also die Knoten eines Patienten mit Kanten verbunden. Zunächst werden alle Knoten einer Messart chronologisch miteinander verbunden. In Abbildung 11 wird dies durch die erste Zeile der Verarbeitung dargestellt („verbinde chronologisch“). In der zweiten Zeile („verbinde normale und anormale Knoten chronologisch“) wird dargestellt, dass Knoten gleicher Messwertart zusätzlich in normale Messwerte und krankhafte bzw. anormale Messwerte unterschieden werden. Innerhalb jeder Messwertart werden also in diesem Schritt alle normalen Knoten untereinander chronologisch verbunden und ebenso alle anormalen Knoten der Messwertart chronologisch verbunden. Im letzten Schritt des Multithreading wird der Startknoten jeder Messwertart nun noch mit dem Patientenknoten verbunden. Die Beschriftung dieser Kante ist „START“.

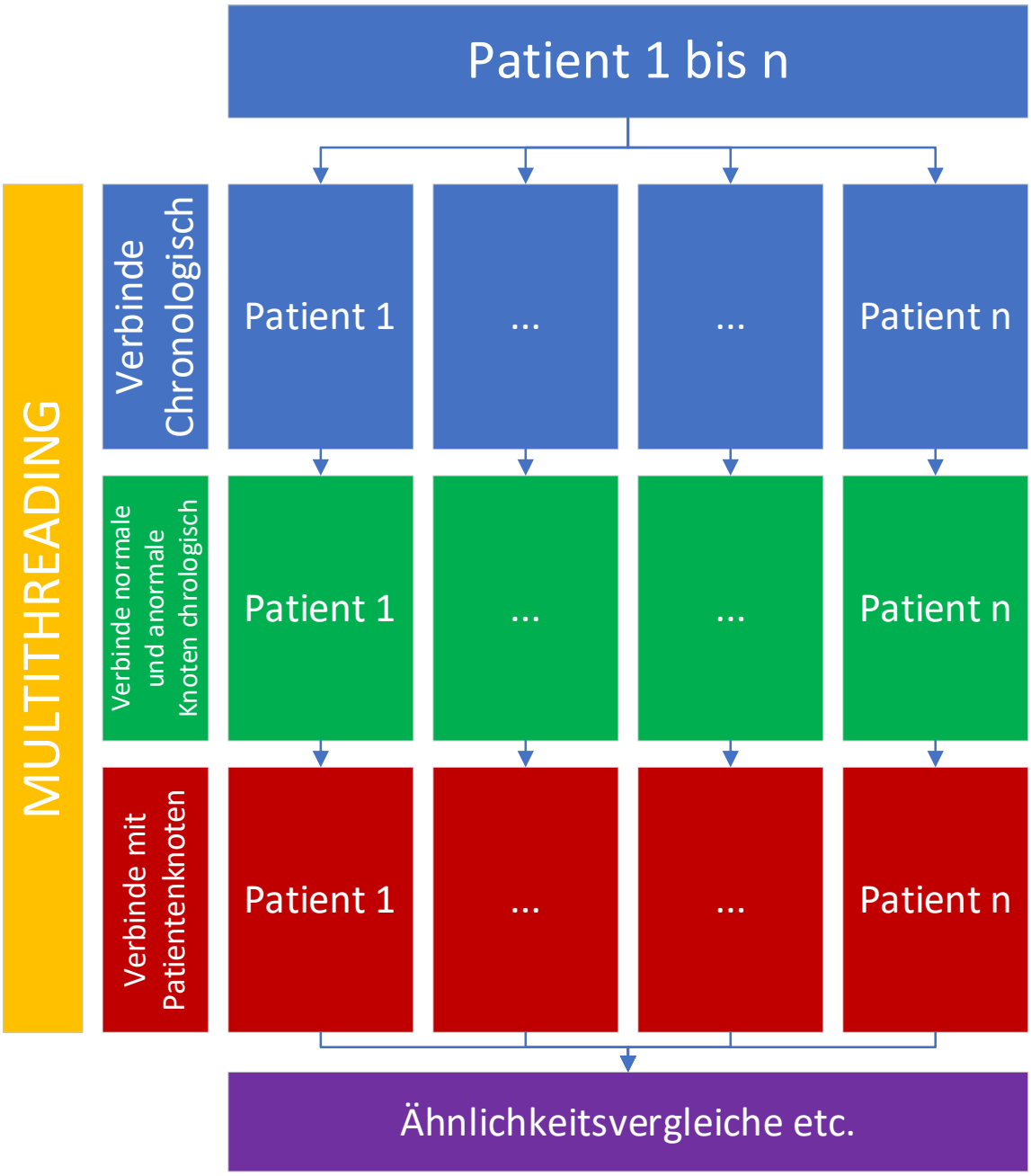

Abbildung 11: Schematische Darstellung des Multithreading beim Erzeugen der Kanten beim Einlesen der Chartevents in die Grafdatenbank. Zunächst werden alle Knoten pro Patient chronologisch miteinander verbunden. Dieser Schritt wird für die Patienten gleichzeitig durchgeführt, begrenzt durch die Anzahl der zur Verfügung stehenden Threads. Anschließend werden die einzelnen Knotenarten innerhalb eines Patienten miteinander verbunden, dabei werden unterschiedliche Zeitstrahlen für normale und anormale Knoten erstellt. Zuletzt werden die zeitlich ersten Knoten jeder Knotenart mit dem Patientenknoten verbunden.

Tabelle 14: Algorithmus zum Einlesen der Chartevent-Patientendaten

Algorithmus 5: Patientendaten einlesen

```
Array listPatienten = ermittle_vorhandene_patienten(vonPatientenNr, BisPatientenNr);
for all i in listPatienten do
 new Thread
 {
  Array listChartevents = ermittle_chartevents_pro_patient(i);
  for all j in listChartevents do
    bool lNormal = ermittle_ob_chartevent_normal_oder_anormal();
    listNodeArt.Add(eindeutigeKnotenArt);
  end for
  for all k in listNodeArt do
     EventItem = listNodeArt[k].cItemID;
     if(ermittle_ob_event_item_id_schonmal_bearbeitet_wurde() == false)
     {
       listKnoten = erzeuge_liste_aller_knoten_pro_patient_und_eventitem(i, EventItem);
       listKnoten = sortiere_list_knoten_nach_min_from_first_charttime(listKnoten);

verbinde_alle_knoten_mit_dem_jeweils_nachfolgenden_knoten_der_liste(listKnoten);
     }
   end for
 }
end for
```

Um gleichzeitig auf die MIMIC-III-Datenbank und die Grafstruktur in nur ein und derselben Abfrage zugreifen zu können, wurde anschließend versucht, die Datenstruktur aus der Grafdatenbank in die PostgreSQL-Datenbank zu übertragen. Dazu wurden die Patientendaten per Desktopanwendung und Grafclient (Idziaszek et al. 2016) für .NET von neo4j in CSV-Dateien exportiert. In diesen CSV-Dateien wurden direkt die Knoten dargestellt (jede Zeile ist ein

Knoten) und durch gesonderte CSV-Tabellen werden die Kanten dargestellt. Bei entsprechender Aufbereitung der CSV-Dateien konnten diese über den Bulkimport von neo4j in vergleichsweiser sehr kurzer Zeit eingelesen werden. Dieser Bulkimport gewährleistet einen überprüfungsfreien und dadurch sehr schnellen Import der MIMICIII-Daten in die Grafdatenbank neo4j. Die Daten, die hier eingelesen werden, müssen zuvor in einem strukturierten CSV-Format aufbereitet werden. Diese aufbereiteten CSV-Dateien ähneln stark dem relationalen Datenbankmodell und konnten deshalb als Grundlage für die Übertragung der Grafdatenbankstruktur in die relationale PostgreSQL-Datenbank helfen. Die relationale Struktur einer Grafdatenbank ist vor allem dadurch definiert, dass für Knotenarten differenzierte Datenbanktabellen erzeugt werden sollen. Außerdem, und das ist der Hauptunterschied einer Grafdatenbank zu einer standardmäßigen relationalen Datenbank, werden für die Verbindungen zwischen den einzelnen Knoten extra Datenbanktabellen angelegt. In diesen Datenbanktabellen werden also die Relationen zwischen den Knoten fest eingespeichert. Bei der Datenmenge, die aus der MIMIC-III-Datenbank resultiert, wirkt sich dieser Wechsel auf eine PostgreSQL-Datenbank nur minimal auf die Geschwindigkeit aus. Durch Setzen der entsprechenden Indizes und Parallelberechnung mehrere Knoten konnte diese Geschwindigkeitsdifferenz minimal gehalten werden und ist bei dieser Datenmenge eher theoretischer Natur.

4.4 Auswahl eines Ähnlichkeitsmaßes

Im Folgenden wird die Auswahl eines Ähnlichkeitsmaßes vorgestellt, welches auf die in dieser Arbeit erarbeiteten Patientengrafen als Vergleichsalgorithmus angewendet werden soll und aus deren Vergleich Rückschlüsse auf Diagnose- und Therapieentscheidungen gezogen werden sollen.

Ein Ähnlichkeitsmaß, das die in Kapital 4.2.1 dargestellten Patientengrafen untereinander vergleichen soll, muss bestimmten Anforderungen gerecht werden. Um diese Anforderungen zu erfassen, ist es zunächst notwendig die kritischen Eigenschaften dieser Grafen zu identifizieren. Da die Grafen Patientendaten beinhalten werden, die theoretisch Daten einer ganzen

Lebensspanne eines Patienten beinhalten könnten, sollten die Algorithmen, die darauf anzuwenden sind, entsprechend skalierbar sein. Die Teilgrafen der Patientengrafen stellen unterschiedliche Messgrößen dar. Diese Teilgrafen können für jeden Patienten unterschiedlich lang sein, im Extremfall besteht ein Teilgraf aus nur einem Element oder sogar aus gar keinem Element. Im anderen Extrem könnte ein Teilgraf mehrere tausend Elemente beinhalten. Dementsprechend muss ein Algorithmus, der zum Vergleich zweier Grafen herangezogen wird, weitestgehend größenunabhängige Vergleiche anstellen können. Der klinische Alltag erfordert außerdem oftmals schnelle Reaktionszeiten, weshalb Algorithmen, die exponentiell steigende Zeiten zur Berechnung des Ähnlichkeitsproblems benötigen, nicht infrage kommen. Aus der Literatur konnten mehrere Algorithmen identifiziert, die als mögliche Kandidaten für Vergleichsalgorithmen für Grafen infrage kommen könnten. Im Folgenden wurden die Algorithmen auf Ihre Anwendbarkeit für diesen speziellen Fall des Vergleichs zweier Patientengrafen hin untersucht und mit anderen Algorithmen verglichen, um so eine Auswahl des besten Kandidaten treffen zu können. In Tabelle 15 wurden alle Kandidaten aufgelistet, die nicht bereits durch den Titel oder Abstract Ihrer Artikel aus der unsystematischen Literaturrecherche aus Kapitel 3.4 ausgeschlossen wurden.

Das Hauptziel des Algorithmus SCAN (Xu et al. 2007) besteht darin Cluster, Verbindungsknoten zwischen Clustern und Ausreißer zu identifizieren. Der Algorithmus setzt dabei darauf die Nachbarschaft eines Knoten zu untersuchen, statt nur die direkten Verbindungen eines Knoten zu untersuchen. Dabei wird von der Prämisse ausgegangen, dass Knoten dann in Cluster gruppiert werden, wenn sie ihre Nachbarn miteinander teilen.

Ähnlich wie bei SCAN besteht das Hauptziel von GSE (Bu et al. 2019) ebenso darin Cluster zu identifizieren. Der Algorithmus eignet sich als typische Clustering-Methode im RNA-Seq-Bereich und dient vor allem der schnellen Verbesserung des Signal-Rausch-Verhältnisses.

Der Algorithmus DP-SCAN (Lin et al. 2019) baut dagegen wiederum auf dem Algorithmus SCAN (Xu et al. 2007) auf. Das Kürzel DP steht dabei für Differential Privacy. Die Erweiterung des SCAN Algorithmus in diesem Beitrag zielt auch im Wesentlichen darauf ab, die differentielle Privatsphäre

unter Beibehaltung der Effizienz der Clusteranalyse zu maximieren. Dabei sollen zwar allgemein Informationen durch den Algorithmus bereitgestellt werden können, jedoch soll gleichzeitig die Privatsphäre einzelner in den Daten enthaltenen Personen gewahrt werden (Dankar und El Emam 2013).

Der Algorithmus NetSimile (Berlingerio et al. 2012) beschreibt dagegen eine Möglichkeit zwei Grafen miteinander zu vergleichen und zwar unabhängig von der Größe der beiden Grafen und skalierbar. Für jeden Grafen wird dabei ein sogenannter Signaturvektor aus verschiedenen Messwerten erzeugt. Die Vektoren der zu vergleichenden Grafen werden anschließend mit der Canberra-Distanz verglichen, woraus dann ein Ähnlichkeitswert resultiert. Je kleiner der Wert, desto ähnlicher sind sich die Grafen.

Ein weiterer Algorithmus, der die Ähnlichkeit von Grafen ermittelt ist FSM (frequent subgraph mining) (Jiang et al. 2013). Der Algorithmus beziehungsweise die Gruppe von Algorithmen konzentriert sich auf die Identifizierung von häufigen Subgrafen in Grafdatensätzen. Dabei werden zwei Hauptforschungsbereiche bei der Entwicklung dieser Algorithmen unterschieden: zum einen die Entwicklung effektiver Mechanismen, um Subgrafkandidaten zu erzeugen ohne Duplikate zu produzieren, zum anderen die Identifizierung des besten Weges Kandidaten-Subgrafen zu verarbeiten, sodass die gewünschten häufigen Subgrafen isoliert werden können und das so effektiv wie möglich (Jiang et al. 2013).

Die Idee hinter der Eigenwert Extraktion (Koutra et al. 2011) und der Anwendung dieser bei der Grafenähnlichkeitsanalyse ist, dass jeder Graf bestimmte Werte zu Durchmesser, Eigenwerte und degree distribution hat (degree distribution beschreibt die Zahl der Verbindungen, die ein Knoten in einem Netzwerk zu anderen Knoten hat). Nachdem diese Werte aus einem Grafen extrahiert wurden, werden sie einer Ähnlichkeitsanalyse unterworfen (Koutra et al. 2011). Auch der Begriff der Eigenwert Extraktion stellt einen Überbegriff für mehrere ähnlich artige Algorithmen dar.

Der Algorithmus GraphSIM (Yang et al. 2020) fokussiert sich bei der Ähnlichkeitsbewertung vor allem auf 2D- und 3D-Bilder. Dabei werden die Bilder in Grafen drei Farbkanälen Rot, Grün und Blau aufgeteilt, die jeweils mehrere repräsentative Bildpunkte beinhalten. Diese Grafen werden dann

durch den Algorithmus mit Grafen eines anderen Bildes verglichen, um automatisiert die Ähnlichkeit zweier Bilder zueinander bestimmen zu können.

Die beiden Algorithmen AMF und AMFP (Shtar et al. 2019) basieren auf einem neuronalen Netzwerk, welches für existierende Grafen abschätzt, welche Kantenverbindungen in der Zukunft zwischen Knoten noch existieren könnten. Dabei wird das neuronale Netzwerk auf über 1.000 Medikamente und über 45.000 Medikamenteninteraktionen trainiert. Die Methode wird mit einem späteren Stand derselben Medikamentendatenbank evaluiert, wonach dann 1.440 Medikamente und 248.146 Interaktionen dokumentiert sind. Die Autoren erhoffen sich daraus Rückschlüsse auf mögliche Medikamenteninteraktionen, die bisher nicht dokumentiert sind (Shtar et al. 2019).

Der Algorithmus *graph edit distance* (GED) (Bunke und Allermann 1983) stellt eine typische Graftransformationsmethode (engl. graph transformation method) dar. Dabei wird die Ähnlichkeit zweier Grafen anhand der minimalen Anzahl der Transformationen berechnet, um einen Grafen in den anderen Grafen zu überführen.

Der Jaccard Index ist eine Berechnungsmethode, die auf einer einfachen Formel basiert. Die Formel lautet $\frac{a+b}{a+b+c}$, , wobei a die Anzahl der Knoten darstellt, die nur im ersten Netzwerk vorhanden sind, b ist die Anzahl der Knoten, die nur im zweiten Netzwerk vorhanden ist und c stellt die Anzahl der Knoten dar, die in beiden Netzwerken vorkommen (López et al. 2019). Je kleiner der Wert, desto ähnlicher sind sich die beiden verglichenen Grafen.

Der A* Algorithmus (Sharma et al. 2012) basiert ebenso wie GED (Bunke und Allermann 1983) auf dem Prinzip der Ermittlung der niedrigsten Transformationskosten zur Überführung eines Grafen in einen anderen, um durch diese Kosten die Ähnlichkeit der beiden Grafen zueinander festzustellen. Der Algorithmus versucht dabei schrittweise einzelne Knoten in ihre Zielknoten zu überführen. Dabei macht sich der Algorithmus die Summe des Abstands eines Quellknoten zum Zielknoten plus einer heuristischen Funktion zunutze, die die minimalen Kosten des aktuellen Knoten n zu seinem Zielknoten schätzt (Sharma et al. 2012). Je größer der Wert summiert über alle Knoten ist, desto unähnlicher sind sich die Grafen.

Der Algorithmus der starPep-Toolbox (Aguilera-Mendoza et al. 2020) ist selbst kein Algorithmus, der eine Ähnlichkeitsberechnung durchführt, sondern dient lediglich der Darstellung von Peptiden und deren Ähnlichkeiten untereinander in einem Ähnlichkeitsgrafen. Dabei werden Ähnlichkeiten als Kanten des Grafen dargestellt, die Peptide selbst als Knoten. Die Ermittlung der Ähnlichkeitswerte ist allerdings nicht Teil des Artikels, sondern lediglich deren Darstellung (Aguilera-Mendoza et al. 2020).

Der Algorithmus RASCAL (Raymond und Willett 2002) berechnet dagegen wiederum Ähnlichkeiten zwischen Grafen, er kann dem Subgraf-Isomorphismus zugeordnet werden. Dabei werden verschiedene bereits früher beschriebene Ähnlichkeitskoeffizienten in den Algorithmus eingebaut. Im Artikel werden Fingerprint-basierte Ähnlichkeitsmessungen mit grafbasierten verglichen, insbesondere mit dem neu entstandenen Algorithmus RASCAL (Raymond und Willett 2002).

Der Algorithmus spgk (shortest-path graph kernel) (Alvarez et al. 2011) ist ein intrinsischer Algorithmus, der exklusiv auf die Gene Ontology ausgelegt ist. Dabei soll die Ähnlichkeit zweier Grafen von Genprodukten auf Basis der Berechnung des kürzesten Pfades ermittelt werden. Diese Ähnlichkeitsberechnung basiert hier auf dem Floyd-Warshall Algorithmus (Alvarez et al. 2011).

Tabelle 15: Auflistung aller recherchierten Algorithmen, die theoretisch für einen Grafenvergleich infrage kommen.

Name	Literatur
SCAN	Xu et al. (2007)
GSE	Bu et al. (2019)
DP-SCAN	Lin et al. (2019)
NetSimile	Berlingerio et al. (2012)
Frequent Subgraph Mining	Jiang et al. (2013)
Eigenvalues Extraction	Koutra et al. (2011)
GraphSIM	Yang et al. (2020) Hu et al. (2019)
AMF/AMFP	Shtar et al. (2019)
GED	López et al. (2019), Bunke und Allermann (1983)
Jaccard index	López et al. (2019), Anderson et al. (2006)
A* Algorithmus	Sharma et al. (2012)
Algorithmus der starPep Toolbox	Aguilera-Mendoza et al. (2020)
RASCAL	Raymond und Willett (2002)
Spgk-Algorithmus	Alvarez et al. (2011)

Im Folgenden wurden die Kandidatenalgorithmen einer Volltextanalyse unterzogen, so konnte die Liste der geeigneten Algorithmen recht schnell verkleinert werden, um zu einer Auswahl gelangen zu können.

So erfüllt der Algorithmus SCAN (Xu et al. 2007) zwar in Titel und Abstract die Vorgaben, um in die Betrachtung mit eingeschlossen zu werden, in der Volltextanalyse wird allerdings schnell klar, dass dieser Algorithmus den Fokus darauf setzt, Ausreißer Knoten und Verbindungsknoten zwischen zwei Clustern als solche zu identifizieren. Dabei geht es weniger um die Ähnlichkeitsberechnung zwischen zwei Grafen, weshalb dieser Algorithmus für die nähere Auswahl der Ähnlichkeitsalgorithmen bereits ausscheidet.

Ähnlich verhält es sich mit dem Algorithmus GSE (Bu et al. 2019). Dieser Algorithmus dient vor allem als Clustering Methode und ist auch weniger als Ähnlichkeitsalgorithmus geeignet und wird deshalb aus der weiteren Betrachtung ausgeschlossen.

Der Algorithmus DP-SCAN baut auf dem SCAN-Algorithmus auf. Die Anpassungen, die Lin et al. (2019) für SCAN bereithalten, schließen allerdings keine expliziten Ähnlichkeitsberechnungen zwischen Grafen mit ein, weshalb dieser Algorithmus ebenfalls aus der weiteren Betrachtung ausgeschlossen wurde.

Der Algorithmus *GraphSIM* (Yang et al. 2020) stellt zwar einen Algorithmus dar, der Grafen auf ihre Ähnlichkeit hin miteinander vergleicht, in dem vorliegenden Artikel wird der Algorithmus aber auf 2D- und 3D-Bilder angewendet und ist rein auf Grafen ausgelegt, deren zentraler Knoten lediglich eine, beziehungsweise einige wenige Ebenen, von Nachbarknoten aufweist. Die in dieser Arbeit zu betrachtenden Grafen sind allerdings weitaus größer und beinhalten teilweise einige tausend Messpunkte. *GraphSIM* wird deshalb ebenfalls nicht in die weitere Betrachtung eingeschlossen.

Dagegen entfernen sich die beiden Algorithmen *AMF* und *AMFP* (Shtar et al. 2019) wiederum von dem Anwendungsziel der Ähnlichkeitsberechnung größerer Grafen. Die beiden Algorithmen basieren auf einem neuronalen Netzwerk, welches nicht ähnliche Grafen berechnet, sondern für existierende Grafen abschätzt, welche Kantenverbindungen in der Zukunft zwischen Knoten noch existieren könnten. Die Autoren erhoffen sich daraus Rückschlüsse auf mögliche Medikamenteninteraktionen, die bisher nicht dokumentiert sind. Für den hier beschriebenen Anwendungsfall sind die beiden Algorithmen damit allerdings ungeeignet.

Der Algorithmus *graph edit distance* (GED) (Bunke und Allermann 1983) stellt eine klassische Graftransformationsmethode (engl. graph transformation method) dar. Dabei wird die Ähnlichkeit zweier Grafen anhand der minimalen Anzahl der Transformationen berechnet, um einen Grafen in den anderen Grafen zu überführen. Die Grafen im Anwendungsfall dieser Arbeit sind allerdings so beschaffen, dass auch Grafen unterschiedlicher Größe trotzdem ähnlich sein können. Der Angleich zweier in der Größe sehr unterschiedlicher Grafen auf

dieselbe Größe würde allerdings einen enormen Transformationsaufwand bedeuten, weshalb dieser Algorithmus an dieser Stelle ebenfalls ausscheidet. Ein hoher Transformationsaufwand würde hier nämlich bedeuten, dass sich die Grafen eher unähnlich sind, obwohl zwei zu betrachtende Patienten vielleicht dieselbe Diagnose haben, aber ein Patient beispielsweise gerade erst eingeliefert wurde, während der andere Patient schon mehrere Wochen Daten in der Intensivstation sammelt und damit einen deutlich größeren Grafen besitzt.

Beim Jaccard Index (Anderson et al. 2006) verhält es sich ähnlich. Dieser Index bezieht sich rein auf die Anzahl der Elemente, die zwei Grafen *a* und *b* gemeinsam haben oder eben nicht gemeinsam haben. Der Index ist also wieder abhängig von der Größe der beiden Grafen und somit ebenfalls für den hier gezeigten Anwendungsfall ungeeignet.

Wie der GED-Algorithmus basiert der A* Algorithmus (Sharma et al. 2012) ebenfalls auf dem Prinzip der niedrigsten Transformationskosten, um einen Grafen in einen anderen zu überführen und ist damit ebenfalls ungeeignet für den hier gezeigten Anwendungsfall.

Auch der Ähnlichkeitsalgorithmus von Aguilera-Mendoza et al. (2020) ist ungeeignet für den hier betrachteten Anwendungsfall. Dieser Algorithmus erzeugt Knoten und die Ähnlichkeit der Knoten wird in den Kanten kodiert. Im hier gezeigten Anwendungsfall werden die Kanten allerdings anders codiert und der Ähnlichkeitsalgorithmus soll die Ähnlichkeit zwischen zwei Grafen berechnen, nicht zwischen zwei Knoten. Insofern kann dieser Algorithmus ebenfalls aus der weiteren Betrachtung ausgeschlossen werden.

Dagegen kann der RASCAL-Algorithmus (Raymond und Willett 2002) dem Subgraf-Isomorphismus zugeordnet werden. Ähnlichkeitsalgorithmen, die darauf beruhen, sind allerdings ungeeignet für den hier betrachteten Anwendungsfall, da diese Algorithmen Grafen als ähnlich betrachten, die möglichst gleich sind, was auch die Größe als Maß der Ähnlichkeit miteinschließt. Ebenso kann das *subgraph isomorphism problem* nur in exponentiell steigender Zeit gelöst werden und damit ist dieser Algorithmus hier nicht geeignet.

Genauso wie GED und A* basiert auch der spgk-Algorithmus (Alvarez et al. 2011) auf dem Prinzip der niedrigsten Transformationskosten und ist damit ebenfalls ungeeignet für die weitere Betrachtung.

Anders als die vorgenannten ausgeschlossenen Algorithmen sind die drei Algorithmen NetSimile, Frequent Subgraph Mining und Eigenvalues Extraction explizit auf die Ähnlichkeitsberechnung von Grafen ausgelegt. Im Folgenden werden diese drei Algorithmen auf Ihre Anwendbarkeit im Zusammenhang mit der in Kapitel 4.2 beschriebenen Grafendarstellung einzelner Patienten und deren Ähnlichkeitsvergleich hin untersucht.

Berlingerio et al. (2012) vergleichen in diesem Zusammenhang ihren Algorithmus *NetSimile* mit den beiden recht verbreiteten Ähnlichkeitsalgorithmen *Frequent Subgraph Mining* (FSM) und *Eigenvalues Extraction* (EIG). Dabei stellten Berlingerio et al. (2012) fest, dass FSM die Eigenschaft der Skalierbarkeit vermissen lässt. Der Grund dafür ist dessen Ursprung beim *subgraph isomorphism problem*. Wie oben bereits beschrieben sind Algorithmen, die auf dem *subgraph isomorphism problem* beruhen, als ungeeignet für den hier gezeigten Anwendungsfall zu betrachten. Dieses Grundproblem der Grafentheorie gilt als Problem, das von theoretischer Seite nur in exponentiell steigender Zeit gelöst werden kann (Ullmann 1976). Aus diesem Grund ist FSM für die Anwendung in dieser Applikation als ungeeignet anzusehen. EIG dagegen ist zwar skalierbar, berechnet die Ähnlichkeiten dafür aber nicht größenunabhängig. Demnach sind sich Netzwerke nach Anwendung mit diesem Algorithmus mit vergleichbarer Größe ähnlicher, als Netzwerke mit unterschiedlicher Größe. Diese Eigenschaft widerspricht den Daten, die in dem hier vorgestellten Patientengrafen gespeichert sein können. Demnach wäre ein Patient, der gerade frisch auf die Intensivstation gekommen ist, einem Patienten mit einer völlig anderen Krankheit, der aber ebenfalls frisch auf die Intensivstation gekommen ist, ähnlicher als einem Patienten mit derselben Krankheit, der aber bereits einen längeren Zeitraum auf der Intensivstation liegt. Außerdem lassen sich bei EIG Ähnlichkeiten untereinander nur sehr schwer vergleichen, da es keinen globalen Höchstwert für Eigenwerte gibt (Berlingerio et al. 2012). Der Algorithmus *NetSimile* (Berlingerio et al. 2012) vereint dagegen die für diese Art von Grafen notwendigen Eigenschaften zum Vergleich dieser Grafen. Der Algorithmus ist skalierbar und damit anwendbar auf verschiedene

Grafengrößen. Außerdem berechnet er die Ähnlichkeiten größenunabhängig, sodass ähnlich große Grafen nicht automatisch einen höheren Ähnlichkeitswert erhalten als Grafen, die unterschiedlich groß sind. Diese Eigenschaften sind zurückzuführen auf die Gestaltung der Logik des Algorithmus. *NetSimile* berechnet verschiedene „Features“ (Berlingerio et al. 2012) eines Grafen, um Kennzahlen für dessen strukturellen Aufbau zu erhalten. Dabei werden diese Features für jeden einzelne Knoten berechnet, beispielsweise sind das die Anzahl der Nachbarn, der Cluster-Koeffizient des Knoten, die durchschnittliche Zahl an Knoten, die zwei Sprünge entfernt sind, et cetera (Berlingerio et al. 2012). Aus diesen Werten werden jeweils fünf weitere statistische Werte berechnet, diese sind beispielsweise der Mittelwert, der Median und die Standardabweichung. Will man nun zwei Grafen vergleichen, so ermittelt man die Canberra-Distanz zwischen dem Mittelwert des einen Features des ersten Grafen mit dem Mittelwert desselben Features des anderen Grafen. Genauso verfährt man bei den vier weiteren Werten und den anderen Features und erhält dadurch die Canberra-Distanz der einzelnen Werte beider Grafen zueinander, welche zusammengenommen nun als Maß für die Ähnlichkeit betrachtet werden können. Je kleiner die Distanz, desto ähnlicher sind sich die Grafen strukturell (Berlingerio et al. 2012). Durch diese Vorgehensweise können strukturell ähnliche Grafen identifiziert werden und zwar ganz unabhängig von der Anzahl der Knoten eines Grafen. Somit ist die Ähnlichkeit auch nicht abhängig von der Anzahl der durchgeführten Messungen an den Patienten.

Eine Ähnlichkeitsanalyse mit Hilfe des NetSimile-Algorithmus kann auf mehrere Arten und Weisen durchgeführt werden. Zum einen kann der gesamte Graf in die Berechnung miteingeschlossen werden, zum anderen können aber auch die Teilgrafen beziehungsweise Zweige für sich in den unterschiedlichen Patientenprofilen miteinander verglichen werden. Dabei werden immer diejenigen Teilgrafen miteinander verglichen, die dieselbe event_item_id haben. Das heißt, dass die Teilgrafen dieselbe Kategorie von Messwerten darstellen müssen, um vergleichbar zu sein. Zum Beispiel wird also der Teilgraf Chlorid von Patient 1 mit dem Teilgraf Chlorid von Patient 2 verglichen. Die Schwierigkeit besteht nun darin einen passenden Wert für die Ähnlichkeit zweier Teilgrafen zu finden, bei denen zwar bei Patient 1 ein Teilgraf vorhanden ist, in Patient 2 aber kein Teilgraf derselben Kategorie vorhanden ist. Der *NetSimile*

Algorithmus ist damit der erste Kandidat, der für die weitere Betrachtung eingeschlossen wird.

Im Rahmen der Literaturrecherche wurden eine Reihe von möglichen Ähnlichkeitsalgorithmen identifiziert, die Grafen miteinander vergleichen können. Die in dieser Arbeit vorgestellte Darstellung von Grafen kann leider nicht durch Algorithmen untersucht werden, die auf dem Prinzip der niedrigsten Transformationskosten oder dem Isomorphismus beruhen. Dadurch fallen bereits die meisten Methoden weg, die sich mit Ähnlichkeiten zwischen Grafen beschäftigen. Aus dieser Literaturrecherche geht nur der Algorithmus *NetSimile* als möglicher Kandidat für einen passsenden Ähnlichkeitsalgorithmus hervor, weshalb im Folgenden auch nur mit dieser Methode weitergearbeitet wurde.

4.5 Anwendung des Ähnlichkeitsmaßes

In diesem Kapitel wird analysiert, wie sich der Algorithmus mit einer Teilmenge von Patienten verhält, um darzustellen, welche Rückschlüsse im Beispiel aus diesen Ergebnissen gezogen werden können. Dabei wurde unter anderem ein Vergleich zweier Mittelwerte durchgeführt, um zu prüfen, ob sich der Algorithmus vom Hintergrundrauschen absetzt. Die genaue Vorgehensweise wurde bereits in Kapitel 3.5 erläutert. Demnach wird überprüft, ob der Algorithmus mehr ähnliche Sepsis-Patienten identifiziert, wenn man Sepsis-Patienten betrachtet, als wenn man Nicht-Sepsis-Patienten betrachtet. Dadurch kann überprüft werden, ob tatsächlich eine Ähnlichkeit berechnet wurde und durch den Unterschied der Mittelwerte kann eventuell sogar abgeschätzt werden, wie gut diese Ähnlichkeit berechnet wird. Für den ersten Test wurden die in Tabelle 16 aufgelisteten Sepsis- und Septikämie-Diagnosen verwendet, um einen Sepsis-Patienten im Sinne dieser Arbeit zu identifizieren.

Zunächst wurde eine Gesamtstichprobenzahl von 1.000 Patienten überprüft, ob eine Tendenz zur Funktionsfähigkeit ermittelt werden kann. Diese Stichprobe enthält $n_1 = 114$ Sepsis-Patienten und $n_2 = 886$ Nicht-Sepsis-Patienten. Bei der Identifizierung der 20 ähnlichsten Patienten wurden für einen Sepsis-Patienten im Schnitt $\bar{Y}_1 = 6{,}307692$ ähnliche Patienten identifiziert, die zuvor

auch tatsächlich die Diagnose Sepsis erhalten hatten. Die Standardabweichung lag hierbei bei 4,06 Bei den 20 ähnlichsten Patienten für einen Nicht-Sepsis-Patienten wurden $\bar{Y}_2 = 3{,}352601$ Patienten identifiziert, die zuvor die Diagnose Sepsis erhalten hatten. Die Standardabweichung lag hier bei 3,08.

Tabelle 16: Sepsis- und Septikämie-Diagnosen nach ICD9 bezogen auf die gesamte MIMIC-III-Datenbank. Diese Diagnosen wurden im ersten Schritt der Prüfung der Funktionalität der Ähnlichkeitsberechnung dazu verwendet einen Sepsis-Patienten im Sinne dieser Arbeit zu identifizieren. Die Spalte # Pat. zeigt das Vorkommen der Diagnosen in Bezug auf den gesamten Datenbestand der MIMIC-III-Datenbank.

ICD9-Code	***Label***	*# Pat.*
003.1	Salmonella septicemia	1
022.3	Anthrax septicemia	0
038.0	Streptococcal septicemia	376
038.10	Staphylococcal septicemia, unspecified	44
038.11	Methicillin susceptible Staphylococcus aureus septicemia	515
038.12	Methicillin resistant Staphylococcus aureus septicemia	118
038.19	Other staphylococcal septicemia	150
038.2	Pneumococcal septicemia [Strept. pneumoniae septicemia]	88
038.3	Sepsis due to anaerobes	110
038.40	Gram-negative sepsis, unspecified	60
038.41	Septicemia due to hemophilus influenzae	4
038.42	Septicemia due to escherichia coli [E. coli]	467
038.43	Pseudomonas Septicemia	127
038.44	Serratia septicemia	27
038.49	Other septicemia due to gram-negative organisms	395
038.8	Other specified septicemias	206
038.9	Unspecified septicemia	3725
054.5	Herpetic septicemia	2
659.30	Septicemia In Labor-Unsp	0
771.81	Bacterial sepsis of newborn, unspecified	225
785.52	Septic shock	2586
995.90	Systemic inflammatory response syndrome, unspecified	40
995.91	Sepsis	1272
995.92	Severe sepsis	3912
995.93	SIRS due to noninfect. process without acute organ dysfunction	64
995.94	SIRS due to noninfect. process with acute organ dysfunction	117
Gesamt		14631

4.5.1 Analyse der Verteilung ähnlicher Patienten

Die beiden Mittelwerte, die in Kapitel 4.5 vorgestellt wurden, zeigen zwar, dass bei Betrachtung eines Sepsis-Patienten mehr Sepsis-Patienten als ähnliche Patienten gefunden werden, als bei der Betrachtung eines Nicht-Sepsis-Patienten, jedoch sagen diese Mittelwerte noch nichts über die Verteilung der Anzahl der Patienten mit und ohne Sepsis unter den zwanzig ähnlichsten Patienten aus. Abbildung 12 betrachtet diesen Aspekt jedoch genauer. Für den ersten Wert bedeutet das beispielsweise, dass 18,97% von allen Nicht-Sepsis-Patienten null Sepsis-Patienten unter deren zwanzig ähnlichsten Patienten vorzuweisen hatten, während beispielsweise beim zwölften Wert bei 15,56% der Sepsis-Patienten aus der Stichprobe unter den zwanzig ähnlichsten Patienten zwölf Sepsis-Patienten waren.

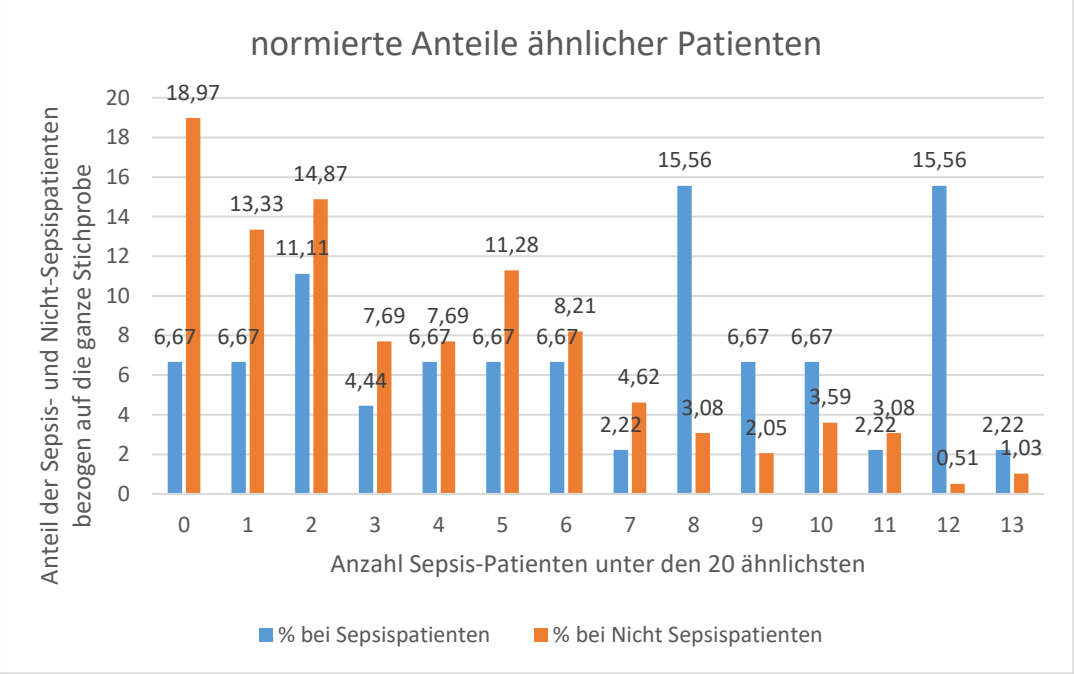

Abbildung 12: auf die Gesamtstichprobe normierte Anteile ähnlicher Patienten. Orange: auf Gesamtzahl an Nicht-Sepsis-Patienten normierter Anteil der Nicht-Sepsis-Patienten, die die auf der X-Achse stehende Anzahl an Sepsis-Patienten unter den 20 ähnlichsten identifiziert haben. Blau: auf Gesamtzahl an Sepsis-Patienten normierter Anteil der Sepsis-Patienten, die die auf der X-Achse stehende Anzahl an Sepsis-Patienten unter den 20 ähnlichsten Patienten identifiziert haben. Die Summen der orangenen und der blauen Anteile ergeben jeweils 100%.

Die Abbildung zeigt demnach, dass für Sepsis-Patienten mehr ähnliche Patienten gefunden werden, die auch tatsächliche Sepsis-Patienten sind, als das

bei Nicht-Sepsis-Patienten der Fall ist. Gerade im Bereich zwischen acht und dreizehn Sepsis-Patienten unter den zwanzig ähnlichsten gibt es eine deutliche Tendenz dahingehend, dass bei Sepsis-Patienten deutlich mehr ähnliche ebenfalls Sepsis-Patienten vorhanden sind als bei Nicht-Sepsis-Patienten. Jedoch muss auch hier betrachtet werden, dass beispielsweise rund 6,25 % der Sepsis-Patienten unter den zwanzig ähnlichsten Patienten gar keinen Sepsis-Patienten identifizieren.

4.5.2 Häufigkeitsanalyse der Diagnosen pro Patient

Bei der Betrachtung der ähnlichsten Patienten muss beachtet werden, dass es nicht den einen ähnlichen oder gleichen Patienten gibt. Die bloße Anzahl an unterschiedlichen Messwertkategorien und Messwerten macht es schlicht unmöglich, einen perfekt passenden Patienten zu finden, da die Messwerte selbst bei derselben Krankheit schon unterschiedlich sein können. Dazu muss außerdem beachtet werden, dass nie von derselben Ausgangssituation ausgegangen werden kann, wie das bei Laborexperimenten der Fall ist. In Laborexperimenten geht man davon aus, dass man zwei Ergebnisse dann vergleichen kann, wenn diese dieselbe Ausgangssituation aufweisen und lediglich in einem oder in voneinander abhängigen Parametern geändert werden (Rieß et al. 2012). Durch Vergleich der beiden Ergebnisse kann auf die Wirkung der geänderten Parameter geschlossen werden. Eine solche Konstellation liegt hier allerdings nicht vor. Die Patientengrafen, die hier miteinander verglichen werden, bestehen jeder einzelne aus unterschiedlichen Patienten. Es kann nicht davon ausgegangen werden, dass zwei Patienten in allen bis auf einer Kategorie Ausprägungen von Messwerten haben. Vielmehr geht es darum, diejenigen Patienten für den aktuell betrachteten Patienten zu identifizieren, die diesem Zustand der maximal möglichen Ähnlichkeit möglichst nahekommen. Dabei ist es möglich, dass für einen Patienten lediglich einige wenige Patienten zu den ähnlichsten Patienten gehören, während für einen anderen Patienten möglicherweise deutlich mehr ähnliche Patienten zu finden sind. Insgesamt geht es bei der Entscheidungsunterstützung im Endergebnis auch nicht so sehr darum, welcher Patient wie ähnlich ist, sondern darum, dem Anwender mögliche

Diagnosen oder Therapien vorzuschlagen, die er selbst vielleicht noch gar nicht in Betracht gezogen hat, die ihn somit also in seiner Entscheidung für die weitere Therapie unterstützen sollen. Durch das System soll ein Mehr an Informationen für den Anwender bereitgestellt werden, sodass er eine fundierte Entscheidung auf Grundlage dieser Datenbasis treffen kann. Im Folgenden wird hierfür eine mögliche Lösung für diese Anforderung dargestellt

Bisher wurde die Anzahl der ähnlichen Patienten mit gleicher Diagnose (Sepsis oder einer Sepsis-ähnlichen Diagnose) und deren Mittelwerten berechnet, um daraus einen Rückschluss auf die Funktionsfähigkeit der Methode zu erlangen. Dabei wurden die Mittelwerte der ähnlichen Sepsis-Patienten und den 20 ähnlichsten Patienten auf Basis eines Sepsis- und eines Nicht-Sepsis-Patienten untersucht. Der Mittelwert entstand dabei aus der Durchführung dieser Untersuchung über alle Patienten. Dadurch kann zwischen dem Grundrauschen eines Nicht-Sepsis-Patienten und dessen ähnlichen Sepsis-Patienten im Vergleich zu den tatsächlichen Sepsis-Patienten und deren ähnlichen Sepsis-Patienten unterschieden werden.

Diese Betrachtungsweise fokussiert sich allerdings sehr auf die Sepsis als Krankheit an sich, wobei die Sepsis eigentlich ein multimorbides Krankheitsbild darstellt (Zador et al. 2019). Das heißt sie kommt sehr häufig in Zusammenhang mit anderen Erkrankungen vor und könnte deshalb ein sehr unspezifisches Bild zeichnen, was wiederum die Ähnlichkeitsmessung stark beeinflussen kann. Aus diesem Grund soll im Folgenden die Betrachtungsweise von der Diagnose Sepsis weg in Richtung der Betrachtung aller Diagnosen gelenkt werden. Dabei soll auch hier zunächst vom Anwendungsfall abstrahiert werden. Der Anwendungsfall kann so beschrieben werden, dass ein neuer Patient untersucht wird, dessen Diagnose unbekannt ist. Durch einen Ähnlichkeitsvergleich und anhand von wenigen Daten des neuen Patienten soll ein möglichst ähnlicher Patient gefunden werden und aus diesem Fund sollen Rückschlüsse auf die Diagnose des neuen Patienten geschlossen werden können.

Dieser Anwendungsfall stellt zwar das letztendliche Ziel der hier dargestellten Patientenvergleiche dar, ist aber für den Programmtest ohne Experten, die das gefundene Krankheitsbild des neuen Patienten identifizieren, ungeeignet. Um beispielhaft die Ergebnisse des Vergleichs darzustellen, sind vorhandene

Patienten und deren Diagnosen deutlich besser geeignet, da so anhand der gestellten Diagnosen für den zu untersuchenden Patienten direkt Rückschlüsse auf den Ähnlichkeitsgehalt der vorgeschlagenen Diagnosen gezogen werden können. Diese Diagnosen wurden vorab bereits von Experten erstellt und damit bietet sich eine einfache Möglichkeit, die ermittelten Ähnlichkeitswerte auf Plausibilität zu prüfen. Dabei muss noch erwähnt werden, dass die zuvor gestellten Diagnosen, die in der MIMIC-III-Datenbank gespeichert sind, keinen Einfluss auf die Ähnlichkeitsberechnung haben. Die Ähnlichkeitswerte der Patientengrafen werden ausschließlich aus den Messwertkategorien der Labordaten berechnet. Die Diagnosen beeinflussen also nicht die Ergebnisse und können deshalb zur Überprüfung herangezogen werden. Es ist hier also explizit nicht der Fall, dass die Diagnosedaten aus der MIMIC-III-Datenbank Einfluss auf die Ähnlichkeitsberechnung als solche haben. Deshalb können diese Diagnosedaten jetzt zur Prüfung der Ähnlichkeitswerte herangezogen werden.

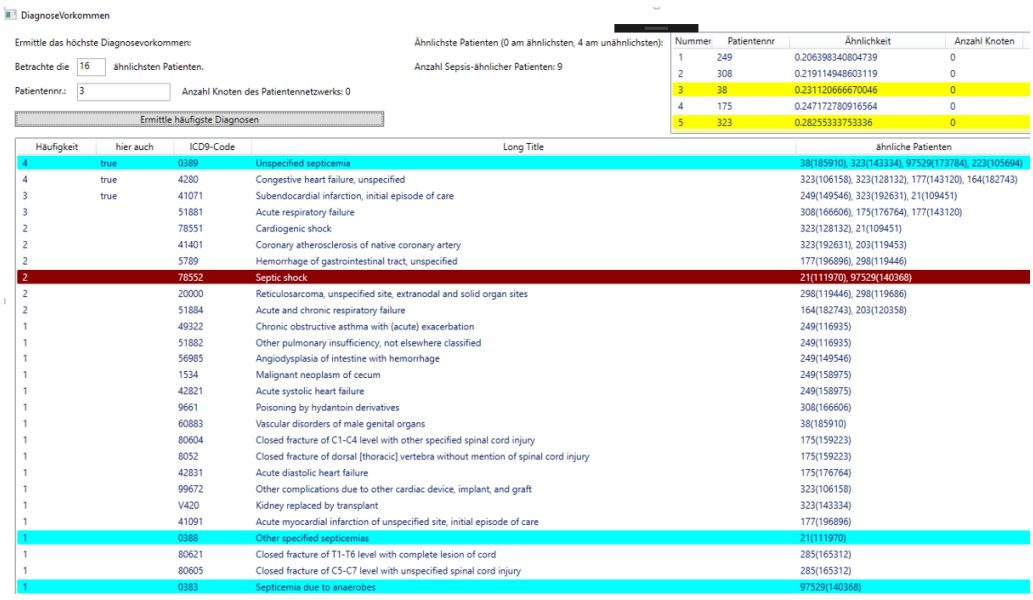

Abbildung 13: Potentieller Anwenderdialog : Ein möglicher Anwenderdialog, der rechts oben die ähnlichsten Patienten zeigt (abhängig von der angegebenen Zahl links) und markiert alle nach Diagnose identifizierten Sepsis-Patienten gelb. Im Feld „Patientennr." wird die zu untersuchende Patientennummer eingetragen. In der Liste unten erscheinen dann die Diagnosen der ähnlichsten Patienten nach Häufigkeit sortiert. In der zweiten Spalte wird markiert, welche der Diagnosen für den aktuell zu untersuchenden Patienten zutreffen. In der vierten Spalte wird dargestellt, welche Patienten in welchem Krankenhausaufenthalt diese Diagnose erhalten hatten.

In Abbildung 13 wird ein möglicher Anwenderdialog zur Entscheidungsunterstützung bei Diagnosen dargestellt, der im Rahmen dieser

Arbeit erstellt wurde. Dabei ist es möglich die Patientennummer anzugeben, die untersucht werden soll, sowie die Anzahl der ähnlichsten Patienten (x), die im Dialog dargestellt werden sollen. In der Liste rechts werden die x ähnlichsten Patienten dargestellt. Gelb markierte Patienten wurden von Experten vorab als Sepsis-Patienten eingestuft, dabei wurden alle Patienten als Sepsis-Patienten festgelegt, für die eine ICD-9-Diagnose aus Tabelle 16 in der MIMIC-III-Datenbank gespeichert war. In der Tabelle darunter werden die Diagnosen der ähnlichsten Patienten nach Häufigkeit sortiert dargestellt. In der letzten Spalte werden die Patienten unter den x ähnlichsten Patienten aufgelistet, die ebenfalls diese Diagnose erhalten hatten. In Spalte 2 werden die Diagnosen markiert, die beim untersuchten Patienten ebenso auftraten. Die Diagnosen des untersuchten Patienten gingen allerdings nicht in die Häufigkeitsberechnung mit ein. An dieser Stelle kann also für Patient 3 in Abbildung 13 die Aussage getroffen werden, dass die Diagnosen, die unter den 16 ähnlichsten Patienten am häufigsten auftraten, genau die Diagnosen waren, die der Patient 3 ebenso erhalten hatte. Dabei gingen in die Ähnlichkeitsberechnung keinerlei Diagnoseschlüssel der ähnlichen Patienten mit ein und auch keine Diagnosen des Patienten 3. Dieses Ergebnis konnte zunächst stichprobenartig bei verschiedenen Patienten in ähnlicher Art und Weise bestätigt werden, deshalb wurde dieser mögliche Zusammenhang im Anschluss genauer untersucht.

Für eine Untersuchung über alle 1.000 Patienten, für die zuvor eine Ähnlichkeitsberechnung durchgeführt wurde, wurde die Software zunächst so umgeschrieben, dass sie alle verfügbaren Patienten aus der Datenbanktabelle lädt, in der die Ähnlichkeiten der Patientengrafen gespeichert sind. Anschließend werden die 20 ähnlichsten Patienten zu den jeweiligen untersuchten Patientennummern geladen. Zu jedem der ähnlichen Patienten werden alle Diagnosen aus den Datenbanktabellen herangezogen und in Bezug zum untersuchten Patienten wird dadurch pro Diagnose die Häufigkeit der Diagnose unter den 20 ähnlichsten Patienten aufsummiert. Außerdem wird für jede der aufgelisteten Diagnosen zum jeweils untersuchten Patienten ermittelt, ob dieser Patient die jeweilige Diagnose ebenfalls erhalten hat. Ist das der Fall, wird für diese Diagnose beim untersuchten Patienten ein entsprechender Marker gesetzt. Die Daten wurden anschließend in eine CSV-Datei exportiert, deren Spalten in Tabelle 17 aufgelistet wurden.

Tabelle 17: Spalten der CSV-Exportdatei der 20 ähnlichsten Patienten eines untersuchten Patienten

Spaltenname	***Beschreibung***
Patientennummer untersuchter Patient	Beinhaltet die Patientennummer des aktuell untersuchten Patienten
ICD9-Code	ICD9-Diagnoseschlüssel ohne Trennzeichen
Marker für Diagnose beim aktuell untersuchten Patienten	true, wenn der aktuell untersuchte Patient die Diagnose ebenfalls erhalten hatte, ansonsten leerer Eintrag
Long label	Die Langbeschreibung des ICD9-Diagnoseschlüssel
Patienten und deren Krankenhausaufenthalte	Patienten kommagetrennt und die zugehörigen Krankenhausaufenthaltsnummern in Klammern
Häufigkeit	Häufigkeit wie oft die Diagnose bei den ähnlichen Patienten des untersuchten Patienten auftrat.

Zunächst wurde für 1.000 untersuchte Patienten ermittelt, wie viele der Patienten in Spalte 3 einen Marker in mindestens einer der gefundenen Diagnosen aufwiesen. Tatsächlich wurde bei 91,9 % der Patienten dieser Marker bei mindestens einer der gefundenen Diagnosen gesetzt. Demnach wurde bei den meisten untersuchten Patienten mindestens eine Diagnose gefunden, die die ähnlichen Patienten auch hatten. Eine kritische Auseinandersetzung hierzu findet sich in Kapitel 5.1.

Die reine Anzahl an Patienten, die eine passende Diagnose hatten, erscheint allerdings noch nicht sonderlich aussagekräftig. Deshalb wurde im Weiteren noch die Häufigkeit der Diagnosen untersucht. Die Diagnosen wurden ja bereits nach ihrer Häufigkeit sortiert. Dabei wurden alle Diagnosen der 20 ähnlichsten Patienten des untersuchten Patienten zusammengefasst und die Häufigkeit der vorkommenden Diagnosen unter den 20 Patienten wurde ermittelt. Die am häufigsten vorkommenden Diagnosen erhielten die Platzierung 1, Diagnosen mit

derselben Häufigkeit erhielten dieselbe Platzierungsnummer, Diagnosen mit weniger Häufigkeit wurden entsprechend in der Platzierung abgestuft. Dabei konnten vereinzelt bis zu 15 Platzierungen und entsprechend nochmal mehr Diagnosevorschläge für jeden untersuchten Patienten ermittelt werden. Auch deshalb, weil viele Diagnosen nur einmal unter den 20 Patienten vorkamen. Anschließend wurde ermittelt, in welcher Platzierungsstufe wie oft für den untersuchten Patienten eine Diagnose stand, die er selbst auch erhalten hatte. Tabelle 18 zeigt diese Daten. In Spalte 1 werden die Platzierungen aufgelistet, sortiert nach dem Quotienten in Spalte 4. Dieser Quotient setzt sich zusammen aus der Zahl der Diagnosen, die der untersuchte Patient ebenso in dieser Platzierung hatte, im Verhältnis zu dem Gesamtvorkommen der Platzierung unter der Gesamtzahl der Patienten. Die Gesamtzahl der Platzierung ist deshalb teilweise deutlich höher als die Gesamtzahl der untersuchten Patienten, weil Platzierungen wie oben beschrieben nach der Häufigkeit der Diagnosen unter den 20 ähnlichsten Patienten vorgenommen werden, wobei Diagnosen mit derselben Häufigkeit diese Platzierungsnummer erhalten haben.

Tabelle 18: Verhältnis der Zahl der Diagnosen, die beim untersuchten Patienten in einer bestimmten Platzierung vorkamen zur Zahl des Gesamtvorkommens der Platzierung.

Platzierung	Marker bei untersuchtem Patienten	Gesamtvorkommen	Quotient Zahl der Markierungen durch Gesamtvorkommen
1	476	1537	0,30969421
2	396	1396	0,28366762
3	367	2250	0,16311111
14	14	202	0,06930693
4	457	7132	0,0640774
5	570	15915	0,03581527
8	372	11415	0,0325887
9	358	11948	0,02996317
13	94	3254	0,02888752
11	382	13646	0,02799355
10	382	14566	0,02622546
6	561	23169	0,02421339
7	438	18537	0,02362842
12	259	13783	0,01879126
15	9	783	0,01149425

Wie aus der Tabelle hervorgeht, wurden beispielsweise unter den drei vordersten Platzierungen im Verhältnis zu deren Gesamtvorkommen die meisten Marker gesetzt. Abbildung 14 verdeutlicht das noch einmal anschaulich. Aus der Abbildung lässt sich herauslesen, dass die ersten drei Diagnosen, die durch die Ähnlichkeitsanalyse gefunden wurden, in vielen Fällen auch Diagnosen waren, die die untersuchten Patienten selbst erhalten hatten. In Zahlen ausgedrückt, wurden in 67,9% der Fälle diese Diagnosen unter den ersten drei Platzierungen gefunden.

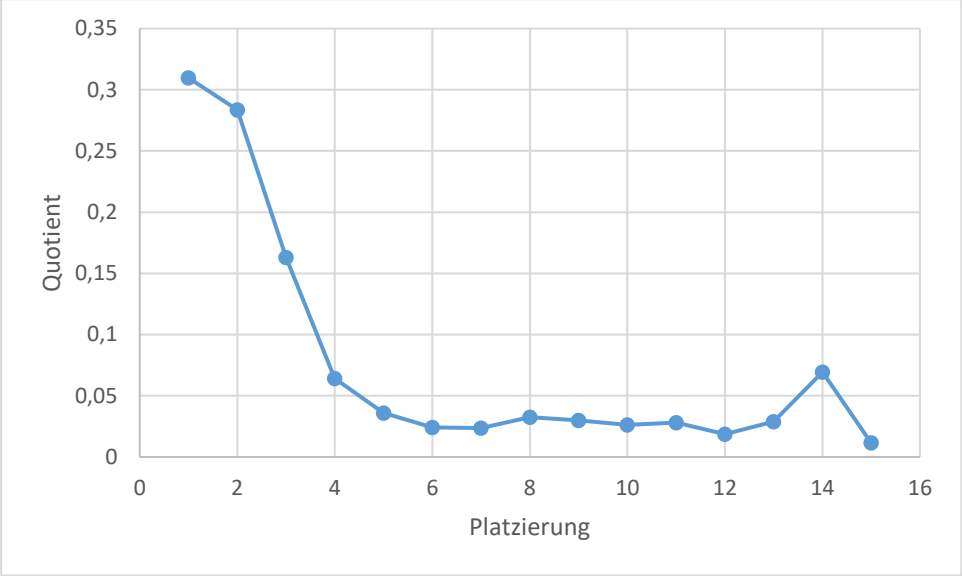

Abbildung 14: Diagnosenplatzierungen gegen den Quotienten der Platzierungen im Gesamtvorkommen: Der Quotient aus Tabelle 18 wurde auf der y-Achse dargestellt, auf der X-Achse die jeweilige Platzierung.

5 Diskussion und Ausblick

5.1 Diskussion der Ergebnisse

Im Rahmen dieser Arbeit wurde eine Methode entwickelt, die als Grundlage für ein Entscheidungsunterstützungssystem betrachtet werden kann und die es ermöglicht, Daten eines Patienten in einen Grafen zu konvertieren, der dann mit Grafen anderer Patienten verglichen werden kann. Der in dieser Methode entwickelte Graf kann dann in ein Entscheidungsunterstützungssystem für Diagnose- und Therapieentscheidungen integriert werden. Bei der Entwicklung dieser Methode entstanden folgende Ergebnisse:

- Eine systematische Literaturstudie, die als Vorarbeit zur eigentlichen Methodenentwicklung die Dissertationsschrift in das Forschungsfeld einbettet.
- Eine Darstellungsform von Patientendaten als Graf
- Eine effiziente Speichermethode zur Speicherung der Grafen in einer geeigneten Datenbank
- Auswahl und Anwendung eines Ähnlichkeitsmaßes zur Darstellung der Möglichkeiten, die durch die entwickelte Methode in einem Entscheidungsunterstützungssystem umgesetzt werden können.

Systematische Literaturstudie

In einer systematischen Literaturstudie (beschrieben in Kapitel 4.1) wurde zunächst untersucht, inwieweit bereits Grafen von Patienten erstellt wurden, die mit der Grafentheorie vereinbar waren und wie diese Grafen dann dazu genutzt wurden, um Patienten miteinander zu vergleichen oder andere Rückschlüsse aus diesen Grafen zu ziehen. Dabei lag das Hauptaugenmerk darauf, dass nur Artikel mit in die Literaturstudie eingeschlossen wurden, die pro Patient einen einzelnen Grafen erzeugten. Bei der Untersuchung konnte außerdem festgestellt werden, dass viele der Artikel das Wort Graf in ganz anderer Weise nutzten, als in grafentheoretischem Zusammenhang. So wurde das Wort Graf sehr häufig als Synonym für Illustration, Bild oder Grafik genutzt. Nur wenige der untersuchten Artikel nutzten einen Grafen im Kontext der Grafentheorie und noch weniger

Autoren nutzten diese Grafen dann so, dass sie damit pro Graf einen individuellen Patienten darstellten. In der Literaturstudie wurden gerade einmal elf Artikel aus fast 400 Artikeln identifiziert, die die Kriterien der Literaturstudie erfüllten. Die Ergebnisse der Literaturstudie ließen keinen Hinweis darauf erkennen, dass die Darstellung von Patienten als Grafen nicht sinnvoll sein könnte. Im Gegenteil, diese elf Artikel zeigten mit ihren sehr unterschiedlichen aber eindrücklichen Ergebnissen, dass Grafdarstellungen von Patienten machbar sind. Aus theoretischer Sicht kann die Modellierung von Patientendaten als Graf sogar einige Vorteile mit sich bringen, speziell in der Analyse dieser Patientendaten, da die Grafentheorie gut etablierte Werkzeuge und Methoden bereithält. Die Nutzung solcher etablierten Algorithmen könnte die Entwicklung von neuen Methoden enorm vereinfachen. Die erste Untersuchungsfrage der Literaturstudie bezog sich deshalb auf mögliche Darstellungen von Patientendaten in Grafen. In den Artikeln wurden verschiedene Möglichkeiten diskutiert, um Knoten und Kanten zu nutzen um Patientendaten der untersuchten Artikel darzustellen. Es wurden vor allem Labordaten als Knoten und temporale Beziehungen als Kanten zwischen den Knoten in den untersuchten Artikeln genutzt. Der Grafentyp, der bei den eingeschlossenen Artikeln am öftesten vorkam, war das *temporal event data mining*, gefolgt von kausalen Netzwerken, Datenbankstrukturansätzen und heterogenem Data-Mining. Der Fokus auf die zeitlichen Zusammenhänge der Knoten zeigt, dass die meisten Daten in diesem Forschungsfeld im Kontext temporaler Beziehungen untersucht wurden, aber die geringe Anzahl an Artikeln, die durch die Literaturstudie eingeschlossen wurden, zeigt außerdem, dass hier noch viel Potenzial für weitere Analysen vorhanden zu sein scheint. Mit der zweiten Forschungsfrage, die in der Literaturstudie untersucht wurde, sollte ein Überblick über alle verschiedenen Mechanismen zur Verarbeitung der untersuchten Patientengrafen gewonnen werden. Die Konzentration von vielen der elf Artikel auf die nahe Vergangenheit könnte anzeigen, dass dies ein recht junges Forschungsfeld ist, welches sich in den nächsten Jahren ausweiten könnte. Zum gegenwärtigen Zeitpunkt ist die Zahl der Artikel, die mit dem Forschungsfeld verbunden sind, jedoch zu niedrig um eine klare Aussage über die weitere Entwicklung des Forschungsfeldes treffen zu können. Weiterhin nutzten lediglich zwei der untersuchten Artikel einen Verarbeitungsmechanismus, nachdem die Patientendaten in einem Grafen

zusammengefasst wurden. Im Gegensatz dazu nutzten die Autoren der restlichen neun Artikel die Grafen nur, um die Patientendaten im Grafen darzustellen, während fünf der eingeschlossenen elf Artikel die Grafen außerdem dazu verwendeten, um die Daten in einer spezifischen Form zu speichern. Diese geringe Zahl von Artikeln, die Grafen nutzten, um individuelle Patienten darzustellen und die noch geringere Zahl von Artikeln, die diese Grafen auch noch weiterverarbeiten, ließen aus der systematischen Literaturstudie verschiedene Fragen aufkommen, die einer weiteren Untersuchung bedurften: Ist es wirklich sinnvoll Patientendaten in Grafen darzustellen und diese weiterzuverarbeiten? Oder gibt es einen bislang unbekannten Grund dafür, warum dies bisher noch nicht allzu oft durchgeführt wurde? Was ist außerdem der beste Weg um die Plausibilität eines solchen Systems testen zu können? Diese Fragen wurden im Kontext diese Arbeit aufgegriffen und beantwortet.

Darstellung der Patientendaten als Graf

In dieser Arbeit wurde eine effiziente Möglichkeit erarbeitet, um Patienten als Grafen darzustellen und die Grafen so aufzubauen, dass sie untereinander vergleichbar werden. Der Sinn und Zweck ist langfristig die Etablierung eines Entscheidungsunterstützungssystems, mit dessen Hilfe Patienten durch Einsatz von CBR miteinander verglichen werden können. Ein Anwendungsfall könnte sein, dass die Daten eines Patienten, der mit unklarer Diagnose in ein Krankenhaus eingeliefert wird, durch die Anwendung in einen Grafen übersetzt werden. Dieser Graf wird dann mit Grafen bereits erfasster Patienten mit erfasster Diagnose über einen Algorithmus verglichen. Die ähnlichsten Patienten können dann Rückschlüsse auf Diagnose und Therapie des Patienten mit unklarer Diagnose liefern. Diese Arbeit hat sich mit der Frage beschäftigt, wie Patientendaten am besten in Grafen dargestellt werden können, um dem Ziel eines grafbasierten EUS einen Schritt näher zu kommen. Diese konkrete in dieser Arbeit entwickelte Darstellungsform erfüllt somit Ziel 1 aus dem Kapitel 1.3.

In Kapitel 4.1.1 wurde eine solche Netzwerkstruktur vorgestellt, die es ermöglicht Patientendaten als Grafen darzustellen. Durch eine vereinheitlichte Darstellung der Patienten als Netzwerk können die Patientennetzwerke

verglichen werden. Zu beachten galt es dabei, dass der Umfang der vorhandenen Daten von Patient zu Patient unterschiedlich sein kann. Es gibt Patienten, über die einige wenige Datenpunkte gespeichert wurden. Es gibt aber auch Patienten, für die tausende von Datenpunkten in der MIMIC-III-Datenbank vorhanden sind. Um trotz der weit differierenden Datenmengen Patientenvergleiche durchführen zu können, ist ein Vergleichsalgorithmus notwendig, mit dem man Netzwerke unabhängig von Ihrer Größe vergleichen kann. Ein Beispiel für einen solchen Algorithmus stellt NetSimile dar (Kapitel 2.6). NetSimile ermittelt Kennzahlen für einen Grafen, die rein auf der Struktur des Grafen basieren. Dementsprechend muss die Struktur des Grafen so gestaltet sein, dass er ebenfalls die für einen Vergleich notwendigen Informationen enthält. Durch die Kategorisierung der Messwerte kann die vierte Anforderung erfüllt werden, nämlich dass eine Netzwerkstruktur der Daten etabliert werden muss, um die Grafen untereinander vergleichbar zu machen. Aus medizinischer Sicht ist es zum Beispiel notwendig, die Werte der Datenpunkte interpretieren zu können, beziehungsweise bewerten zu können, ob ein Datenpunkt normale oder anormale Werte enthält und daraus Schlussfolgerungen zu ziehen. Um diese Wertigkeit mit in die Betrachtung einzuschließen, wurden die Datenpunkte hinsichtlich ihrer Normalität bewertet. Um diese Bewertung auf die Netzwerkstruktur zu übernehmen, wurde sie als Teil der Verzweigungskomplexität kodiert. So wurden alle normalen Werte einer Messwertkategorie in chronologische Reihenfolge gesetzt, genauso wie die anormalen Werte einer einzelnen Messwertkategorie in chronologische Reihenfolge gesetzt wurden. Damit tatsächlich eine Netzwerkstruktur mit Verzweigungen entsteht und nicht nur ein Graf in Sternform mit einem zentralen Punkt, von dem unterschiedliche Äste ganz ohne Verzweigung weg gehen, mussten zusätzliche Verzweigungen eingebaut werden. Diese Anforderung wurde dadurch umgesetzt, dass die getrennt voneinander, chronologisch verbundenen Messwerte jeder Kategorie noch einmal chronologisch im Gesamten verbunden wurden. Dadurch wird jeder Datenpunkt einer Messwertkategorie zusätzlich in eine gemeinsame chronologische Abfolge eingeordnet. Durch die zusätzliche chronologische Verbindung werden normale und anormale Datenpunkte miteinander verbunden und es entsteht ein verbundenes Netzwerk, statt eines rein sternförmig aufgebauten Netzwerkes.

Die Originalmesswerte wurden aber ebenso beibehalten. Diese Flexibilität ermöglicht den Einsatz weiterer Grafenalgorithmen aus der Grafentheorie, da der Einsatz von anderen Algorithmen auch zu der jeweiligen Grafstruktur passen muss. Die Entwicklung einer Darstellungsform und der Einsatz eines Ähnlichkeitsmaßes sind meist sehr eng miteinander verwoben, da bei bereits existierender Darstellungsform und existierendem Ähnlichkeitsmaß mindestens eines dieser beiden angepasst werden muss, dass das Ähnlichkeitsmaß die richtigen Daten erhält, um überhaupt einen aussagekräftigen Vergleich zu erstellen. Die doppelte chronologische Verbindung eröffnet dagegen die Möglichkeit, dass jede Krankheit ihre spezifische Netzwerkstruktur erzeugen kann. So kann die eine Krankheit beispielsweise lange Phasen von normalen, die andere lange Phasen von anormalen Werten einer Kategorie beinhalten, während eine andere Krankheit sprunghaft zwischen normalen und anormalen Werten einer Kategorie wechseln könnte. So könnte die Netzwerkstruktur einer Krankheit eine Art Fingerabdruck für die Identifizierung der Krankheit bei anderen Patienten darstellen. Da Netzwerke unterschiedlicher Patienten auch unterschiedliche Strukturen aufweisen, selbst bei derselben Krankheit, sind nur Ähnlichkeitsanalysen - keine hundertprozentige Identifizierung - möglich, um festzustellen, welche Diagnose am besten zu den Daten des Patienten passt. Wie eindeutig diese Strukturen für einzelne Krankheiten tatsächlich sind, wurde durch diese Arbeit am Beispiel der Sepsis mit untersucht. Gerade in diesem Fall treten neben der Sepsis nämlich oftmals eine Vielzahl verschiedener Begleiterkrankungen und -symptome auf. Auch für Mediziner ist eine Sepsis nicht immer leicht zu erkennen, sondern wird oftmals erst erkannt, wenn die Krankheit bereits so weit fortgeschritten ist, dass die Erfolgsaussichten bereits drastisch gesunken sind. Um solche Situationen möglichst zu vermeiden, gibt es in der Wissenschaft unterschiedliche Ansätze. Einer dieser Ansätze ist der Einsatz von Entscheidungsunterstützungssystemen. Die Untersuchung der Grafendarstellung durch Auswahl eines Ähnlichkeitsmaß in Kapitel 4.4 zeigt deutlich, dass durch ein solches Verfahren in vielen Fällen eben nicht eindeutig auf eine bestimmte Krankheit zurückgeschlossen werden kann. Allerdings kann das Verfahren dazu verwendet werden, um Häufungen in bestimmten Diagnosen zu erkennen. So gehen beispielsweise mit einer Sepsis häufig eine Reihe anderer Diagnosen einher. Manche sind mehr andere weniger spezifisch für Sepsis und

können aber auch nur zufällig gleichzeitig auftreten, weil der Patient bereits einen bestimmten medizinischen Hintergrund hat. Trotzdem werden durch das Anzeigen der Diagnosen von zum Beispiel den zehn ähnlichsten Patienten für einen untersuchten Patienten dem Entscheider eine eingegrenzte Zahl von Diagnosen präsentiert, die mit diesen zehn ähnlichsten Patienten assoziiert sind. Im Idealfall findet der Entscheider dann bei den zehn ähnlichsten Patienten zehn Mal die Krankheit, die der untersuchte Patient auch tatsächlich hat. Doch selbst wenn dieser Idealfall nicht eintritt, so kann eine mögliche Anwendung dennoch so aussehen, dass dem Entscheider Diagnosen vorgeschlagen werden, die oftmals mit der Zieldiagnose assoziiert sind. Allein diese begrenzte Zahl der vorgeschlagenen Diagnosen kann bereits entscheidenden Einfluss auf die weiteren Untersuchungen und Therapien des Entscheiders haben und ihn in die richtige Richtung lenken. Letztendlich entscheidet er aber selbst, ob er den vorgeschlagenen Diagnosen folgen möchte oder nicht. Auf diese Weise generiert er aber möglicherweise Diagnosen, die neue Ideen liefern und zuvor möglicherweise schlicht noch nicht bedacht wurden.

Speicherverfahren

Diese Anwendung kann, ausgehend von der in dieser Arbeit vorgestellten Darstellungsform der Grafen, mit unterschiedlichen Vergleichsalgorithmen umgesetzt werden. Ein Kriterium dafür, dass ein Einsatz verschiedener Vergleichsalgorithmen möglich ist, ist eine effiziente und flexible Speichermethode der Grafen. Die gewählte Speichermethode ermöglicht außerdem eine flexible Anpassung an unterschiedliche Anforderungen möglicher einzusetzender Vergleichsalgorithmen. So wurde gezeigt, dass neben der Klassifikation der Messwerte als normal und anormal auch die originalen Messwerte durch die Wahl der Speichermethode im Grafen hinterlegt werden konnte. Ebenso wurden die Zeitpunkte der Messungen und damit die Abstände zwischen den Messwerten mit im Grafen gespeichert, sodass diese Daten bei Bedarf in einen Vergleichsalgorithmus mit einfließen können. Die relationale Datenbankstruktur wurde in dieser Arbeit so angepasst, dass die Daten theoretisch in eine Grafdatenbank konvertiert werden können. Im konkreten Fall wurde in der Arbeit das DBMS PostgreSQL genutzt. Die Struktur wurde dabei

am Bulk-Import der Grafdatenbank neo4j orientiert. Die Entscheidung für diese beiden Datenbankmodelle und die letztendliche Entscheidung für das relationale Datenbankmodell wurde in Kapitel 4.3 beschrieben. Dabei wurde ausgeführt, dass der vermeintliche Geschwindigkeitsvorteil von Grafdatenbanken bei vielen Daten im speziellen Fall, der in dieser Arbeit beschriebenen Grafenstruktur, eben nicht zum Tragen kommt. Das kann unterschiedliche Gründe haben. Zum einen kann es daran liegen, dass Knoten in diesem speziellen Anwendungsfall nur maximal vier Kanten haben können, während in den meisten Anwendungen der Grafdatenbanken normalerweise Knoten mit einigen Hundert oder Tausend Verbindungen untersucht werden. Zum anderen kann der fehlende Geschwindigkeitsvorteil der Grafdatenbank auch in der Art der Nutzung begründet sein. So wird bei typischen Anwendungen für Grafdatenbanken normalerweise ein bestimmter Knoten gesucht und untersucht, hier geht es also um die Effizienz des Findens, während in der Anwendung in dieser Arbeit eher die Topologie des Gesamtgrafen von entscheidender Bedeutung für die weitere Untersuchung ist. Darüber hinaus ist es ebenso möglich, dass durch die Beschaffenheit des NetSimile-Algorithmus die Geschwindigkeitsvorteile verloren gehen, da dieser Algorithmus jeden einzelnen Knoten im Grafen aufsuchen muss und Werte zu dem Knoten berechnen muss. Diese Berechnungen könnten in einer Matrix möglicherweise besser umgesetzt werden. Möglicherweise hätte auch ein anderes Ähnlichkeitsmaß mehr von der Beschaffenheit der Grafdatenbank profitiert. In dieser Arbeit wurde also ein effizientes Speicherverfahren entwickelt und wurden die Effizienzunterschiede zwischen relationalen und grafenbasierten Datenbanken im Zusammenhang mit den vorliegenden Daten untersucht. Am Ende steht ein effizientes Speicherverfahren, welches, aufgrund von Geschwindigkeits- und Handhabbarkeitsunterschieden abweichend von Ziel 2 in Kapitel 1.3 nicht ein grafenbasiertes, sondern ein relationales Datenbankmodell zugrunde legt. Dieses Speicherverfahren erfüllt somit die Anforderungen, die in Ziel 2 gestellt wurden.

Auswahl und Anwendung eines Ähnlichkeitsmaßes

Der eben erwähnte NetSimile Algorithmus stellt ein mögliches Ähnlichkeitsmaß dar, welches die Grafen aus dieser Arbeit nutzen kann, um die den Grafen zugrundeliegenden Patienten miteinander zu vergleichen. Dieser Algorithmus wurde im Hinblick auf Ziel 3 in Kapitel 1.3 ausgewählt, mit anderen Algorithmen verglichen (Kapitel 4.4) und auf seine Eignung hin untersucht. Dabei musste NetSimile verschiedene spezifische Anforderungen erfüllen, um als geeignet für die Auswertung der hier dargestellten Grafen angesehen zu werden. Zum einen sollten mit Hilfe des Vergleichsalgorithmus größenunabhängige Grafen miteinander vergleichen werden können. Das ist insbesondere deshalb notwendig, weil im klinischen Alltag häufig Patienten mit kurzen Krankenhausaufenthalten und solche mit längeren Aufenthalten vertreten sind. Diejenigen mit längeren Aufenthalten können schnell ein paar tausend Datenpunkte ansammeln, während diejenigen Patienten mit kurzen Aufenthalten meist nur einige wenige Messwerte vorweisen können. Diese Patienten müssen durch das gewählte Ähnlichkeitsmaß aber trotzdem vergleichbar sein. Dementsprechend sollte ein Ähnlichkeitsmaß als zweite Anforderung die Struktur der Grafen als Vergleichsfaktor zugrunde legen. Außerdem muss das Ähnlichkeitsmaß so flexibel sein, dass durch dessen Einsatz Gesamt- und Teilgrafenvergleiche möglich sind. Diesen Anforderungen folgend muss das gesuchte Ähnlichkeitsmaß also ebenso skalierbar sein. Des Weiteren macht es der klinische Alltag erforderlich, dass Vergleichsalgorithmen schnelle Ergebnisse liefern. Demnach kommen Algorithmen, die nur in exponentiell steigender Zeit eine Berechnung des Ähnlichkeitsproblems durchführen können, nicht infrage. Die in Kapitel 3.4.1 vorgestellten Anforderungen bilden die Grundlage zur Auswahl eines passenden Algorithmus. Diese Auswahl wurde in Kapitel 4.4 beschrieben. Dabei wurden mehrere Algorithmen miteinander verglichen, wobei am Ende nur der Algorithmus NetSimile übrigblieb. Dieser wurde anschließend mit den sehr bekannten Algorithmen *Frequent Subgraph Mining* und *Eigenvalues Extraction* mit dem Algorithmus *NetSimile* verglichen. Die Entscheidung fiel letztlich auf den Algorithmus *NetSimile* zur Nutzung als Ähnlichkeitsmaß in Kapitel 4.5, da *NetSimile* in verschiedenen Gesichtspunkten besser abgeschnitten hat, als die beiden anderen Vergleichsalgorithmen. Der Algorithmus *Frequent Subgraph Mining* wurde ausgeschlossen, da er nicht

beziehungsweise wenig skalierbar ist und damit nicht zur Anforderung der Größenunabhängigkeit passt. Des Weiteren liegt dem Algorithmus das *subgraph isomorphism problem* zugrunde, welches nur in exponentieller steigender Zeit lösbar ist, was ebenfalls ein Ausschlusskriterium für diesen Algorithmus darstellt. Die Tatsache, dass das Ähnlichkeitsproblem nur in exponentiell steigender Zeit gelöst werden kann, steht der Anforderung möglichst schneller Ergebnisse im klinischen Alltag entgegen. Gegen den Algorithmus *Eigenvalue Extraction* sprach, dass er zwar skalierbar ist, aber dass er nicht größenunabhängig genutzt werden kann, um Ähnlichkeiten zu berechnen. Dabei ist ebenfalls zu beachten, dass die Eigenwertberechnung keinen globalen Höchstwert besitzt, was die Vergleichbarkeit der Eigenwerte verschiedener Grafen ebenso einschränkt. Demgegenüber besitzt der Algorithmus *NetSimile* die geforderten Eigenschaften der Skalierbarkeit und Größenunabhängigkeit. Der Algorithmus vergleicht Grafen anhand ihrer Struktur und die Berechnung der Ähnlichkeiten geschieht nicht in exponentiell steigender Zeit. Die Auswahl von *NetSimile* nach den Anforderungen eines Vergleichsalgorithmus im Zusammenhang mit der in dieser Arbeit entwickelten Darstellungsform für Grafen erfüllt somit Ziel 3 aus Kapitel 1.3.

Analyse der Verteilung ähnlicher Patienten

Die Analyse der Verteilung ähnlicher Patienten fördert zutage, dass es einen Anteil von insgesamt circa 45 % der Sepsis-Patienten gibt, bei denen unter den 20 ähnlichsten Patienten wiederum zwischen acht und zwölf Sepsis-Patienten gefunden werden. Dagegen werden bei Nicht-Sepsis-Patienten vor allem im Bereich zwischen einem und sechs Sepsis-Patienten unter den 20 ähnlichsten Patienten ermittelt. Diese Tatsache unterstützt die Ergebnisse aus Kapitel 4.5, wonach die Mittelwerte dieser beiden Vergleichsgruppen bereits in diese Richtung zeigen. Andererseits gibt es ebenfalls einen recht hohen Anteil von 6,25 % der Sepsis-Patienten, die nicht einen einzigen Sepsis-Patienten unter deren 20 ähnlichsten Patienten vorweisen können. Das könnte einerseits dafür sprechen, dass der Algorithmus auf jeden Fall noch optimiert werden muss, andererseits kann es aber auch auf die Problematik bei der Analyse von Sepsis als multimorbidem Krankheitsbild hindeuten. Demnach wäre es möglich, dass

die Heterogenität der Sepsis einen Anteil daran hat, dass die Ergebnisse nicht so eindeutig ausfallen, wie sie vielleicht könnten, wenn ein anderes Krankheitsbild betrachtet werden würde. Dieser Sachverhalt könnte in weitergehenden Untersuchungen überprüft werden. Eine Ausweitung der Stichprobe könnte hier ebenfalls für klarere Ergebnisse sorgen. So ist beispielweise nicht klar, warum der Datenpunkt bei 11 von 20 Sepsis-Patienten in Abbildung 12 so niedrige Zahlen für Sepsis-Patienten bei einem untersuchten Sepsis-Patienten liefert. Eine Möglichkeit wäre, dass das zufällig durch die Stichprobe verursacht wurde. Durch eine größere Stichprobe könnte hier in Folgeuntersuchungen Klarheit geschaffen werden.

Häufigkeitsanalyse der Diagnosen pro Patient

Bei der Häufigkeitsanalyse der Diagnosen wurden zunächst von 1000 Patienten die 20 ähnlichsten Patienten ermittelt und deren Diagnosen nach Häufigkeiten jeweils für den untersuchten Patienten aufsummiert. Anschließend wurde ermittelt, wie oft und an welcher Platzierung der Häufigkeiten der Patient dieselbe Diagnose hatte wie seine ähnlichsten Patienten. Zunächst wurde allerdings allgemein ermittelt, wie hoch der Anteil der Patienten ist, die eine oder mehrere Diagnosen identisch mit den Diagnosen der 20 ähnlichsten Patienten der untersuchten Patienten hatten. Im Ergebnis war dies bei 91,9 % der Fall. Diese große Zahl hört sich zunächst so an, als hätte der Ähnlichkeitsalgorithmus gut funktioniert. Man könnte andererseits aber auch argumentieren, dass Patienten auf einer Intensivstation rasch eine große Zahl an Diagnosen ansammeln, wonach die Wahrscheinlichkeit recht klein sein könnte, dass es Patienten gibt, die keine passenden Diagnosen zu ihren 20 ähnlichsten Patienten haben und das allein wegen der bloßen Zahl an Diagnosen pro Patient.

Um diesen Umstand näher zu untersuchen, wurden anschließend Platzierungen für die häufigsten Diagnosen unter den 20 ähnlichsten Patienten vergeben, wobei Diagnosen mit derselben Häufigkeit auch dieselbe Platzierung erhalten haben. Die Ergebnisse zeigen ein relativ eindeutiges Bild, wonach die ersten drei Platzierungen recht häufig Treffer lieferten. Abbildung 14 zeigt, dass diese drei Platzierungen im Vergleich zu den anderen Platzierungen im Verhältnis zum Gesamtvorkommen der Platzierungen in der Stichprobe einen großen Anteil an

den gesetzten Markern haben. Mit anderen Worten: Schaut man sich die drei häufigsten Diagnosen der 20 ähnlichsten Patienten eines untersuchten Patienten genauer an, so hat der Patient in 67,9 % der Fälle eine passende Diagnose unter diesen drei häufigsten Diagnosen. Diese Daten legen die Schlussfolgerung nahe, dass die Ähnlichkeitsanalyse durchaus passable Diagnosen liefert, die zur Entscheidungsunterstützung genutzt werden können. Dabei muss aber immer beachtet werden, dass diese Diagnosen lediglich eine Möglichkeit darstellen, welche der Anwender des EUS in Betracht ziehen könnte. Das System erhebt keinen Anspruch darauf, dass unter den vorgeschlagenen Diagnosen auch tatsächlich die richtige Diagnose enthalten ist. Nun muss man hier allerdings in Betracht ziehen, dass es Diagnosen auf der Intensivstation gibt, die relativ häufig gestellt werden, sodass es möglich ist, dass diese hohe Zahl allein deswegen zustande kommt, weil die Patienten die immer selben Diagnosen gestellt bekommen, wonach die Häufigkeit dieser Diagnosen unter den ähnlichen Patienten und auch unter den untersuchten Patienten immer recht hoch ist. Um diesen Bias ausschließen zu können, bedarf es allerdings weitergehender Untersuchungen.

Zusammenfassung

Die systematische Literaturstudie bildet den Ausgangspunkt dieser Arbeit und stellt die Abgrenzung der Arbeit im Forschungsgebiet dar. Mit der Erarbeitung der Darstellungsform von individuellen Patientendaten in Grafen und der darauf aufbauenden Auswahl eines möglichen passenden Ähnlichkeitsmaßes zur Darstellung möglicher Anwendung sowie mit der Entwicklung eines effizienten und flexiblen Speicherverfahrens konnte mit dieser Arbeit ein wichtiger Grundstein hin zu einem praxisnahen Entscheidungsunterstützungssystem für den klinischen Alltag geschaffen werden. Die Anwendung der Ähnlichkeitsmessung im Zusammenhang mit der Sepsis zeigt die vielfältigen Möglichkeiten, aber gleichzeitig auch noch die momentanen Schwächen, des Systems auf. So konnte gezeigt werden, dass zumindest für die gewählte Stichprobe eine Tendenz hin zur Identifizierung von Sepsis-Patienten durch das System erfolgen konnte. Durch Verbesserungen an verschiedenen Stellen könnte das System so weit optimiert werden, dass Sepsis-Patienten mit noch höherer

Präzision durch das System erkannt werden können, was letztendlich entscheidenden Einfluss auf die Geschwindigkeit der Diagnosestellung haben kann. Damit kann das Entscheidungsunterstützungssystem dazu beitragen eine Sepsis oder andere schwere Krankheitsbilder frühzeitig zu erkennen, sodass rechtzeitig eine Therapie gestartet werden kann und damit könnten die Überlebenschancen eines Patienten letztendlich massiv erhöht werden. Im Folgenden werden mögliche Verbesserungen des Systems dargestellt. So könnte ein anderer Algorithmus gewählt werden, der Algorithmus kann durch veränderte Gewichtungen bei entscheidenden Markern soweit verbessert werden, dass er auf bestimmte Krankheitsbilder besonders sensibel reagiert oder der Algorithmus wird durch Auswahl anderer oder verbesserter Merkmale beziehungsweise NetSimile-Features optimiert. Welche Optimierungen und welche weiteren Möglichkeiten ausgeschöpft werden können, wird im Ausblick in Kapitel 5.3 näher erläutert.

5.2 Diskussion des Vorgehens

Systematische Literaturstudie

Im Folgenden werden die Grenzen der Literaturstudie untersucht. Im Rahmen der Literaturstudie wurden zunächst lediglich vier große Datenbanken nach Artikeln mit entsprechenden Schlüsselwörtern untersucht. In der Literaturstudie wurden lediglich vier Datenbanken für die Literatursuche genutzt (MEDLINE, Web of Science, IEEE Xplore und ACM digital library). Demnach könnten einige Artikel nicht berücksichtigt worden sein, die in anderen Datenbanken indiziert wurden, nicht aber in den vorliegenden vier Datenbanken. Neben der Einschränkung der Anzahl der Datenbanken, könnte es außerdem einige Artikel geben, die zwar in dieses Thema gehören würden, aber durch die spezielle Datenbankanfrage in den vier Datenbanken nicht berücksichtigt wurden. Außerdem spielt der Publikationsbias in dieser Literaturstudie ebenso eine Rolle. So könnte die geringe Zahl von Artikeln, die Grafen zur Darstellung von individuellen Patienten nutzen, aus einer hohen Rate nicht erfolgreicher Untersuchungen in diesem Forschungsfeld resultieren, die dann nicht veröffentlicht wurden.

Außerdem schränkt die Wahl der Suchbegriffe möglicherweise ebenfalls das Suchergebnis so ein, dass Artikel, die eigentlich eingeschlossen hätten werden können, gar nicht berücksichtigt werden, weil sie von vorneherein durch die Wahl der Suchbegriffe ausgeschlossen wurden. Einige weitere Artikel wurden ausgeschlossen, weil sie weder auf Deutsch noch auf Englisch verfasst wurden. Diese Einschränkung könnte weitere passende Artikel aus der Betrachtung ausschließen. Überraschenderweise blieben am Ende des Ausschlussverfahrens lediglich elf Artikel übrig, die zu den Einschlusskriterien passten. Die geringe Zahl an eingeschlossenen Artikeln in der Literaturstudie kann anders herum aber auch so ausgelegt werden, dass die geringe Zahl an Arbeiten, die sich mit diesem Feld beschäftigen, darauf hindeutet, dass dies ein noch recht junges Forschungsfeld ist. In jedem Fall zeigt die Literaturstudie auf, dass in diesem Bereich noch viel Potenzial für weitere Forschungsarbeiten steckt und die eingeschlossenen Artikel zeigen außerdem auch ein inhaltliches Potenzial dieser Methodik auf. Die geringe Anzahl an Artikeln, die durch die Literaturstudie eingeschlossen wurden, erschwerten aber die Identifizierung klarer Tendenzen, welche Teilforschungsgebiete bei Patientengrafen ein gewisses Potenzial bieten und welche nicht.

Darstellung der Patientendaten als Graf

In der vorliegenden Arbeit wurde eine Form der Darstellung der Patientendaten als Grafen gewählt, die eine Umwandlung der Daten derart notwendig macht, dass sie als Graf dargestellt werden können. Dazu wurden zunächst Anforderungen formuliert, die ein so generierter Graf erfüllen sollte. Die in dieser Arbeit entwickelte Darstellungsform ist also von Anforderungen abhängig, die im Zuge dieser Arbeit erarbeitet wurden. Die Darstellungsform könnte durch zu eng gezogene Grenzen bei diesen Anforderungen zu wenig Spielraum für die Darstellungsform bieten und so die möglichen Ergebnisse einschränken. Zu weit gefasste Anforderungen könnten aber allerdings ähnliche Auswirkungen auf die möglichen Ergebnisse haben. Eine daraus resultierende Darstellungsform könnte so unspezifisch sein, dass sich daraus nur relativ uneindeutige Ergebnisse ergeben. Betrachtet man die Möglichkeit ein anderes

Ähnlichkeitsmaß zu nutzen, kann die gewählte Darstellungsform unter Umständen nicht kompatibel zum gewählten Ähnlichkeitsmaß sein. Andererseits wurde die Darstellungsform bewusst offen gestaltet, sodass beispielsweise nicht nur die Bewertung für normale und anormale Werte vorgenommen und gespeichert wurde, sondern auch die Originalwerte. Diese offene Gestaltung ermöglicht eigentlich die Nutzung weiterer Ähnlichkeitsmaße. Trotzdem kann es sein, dass ein anderes gewähltes Ähnlichkeitsmaß nicht zur Darstellungsform passt. Insofern sollte hier beachtet werden, dass die Entwicklung der Darstellungsform und die Auswahl des Ähnlichkeitsmaßes in dieser Arbeit zwar getrennt voneinander untersucht wurden, sich jedoch eine gewisse Verzweigung dieser beiden Teilbereiche nicht vermeiden lässt. Deshalb ist es durchaus möglich, dass beispielsweise die Darstellungsform etwas anders aussehen könnte, wenn man zu dem Ergebnis kommen würde, ein anderes Ähnlichkeitsmaß zu wählen. Um die Darstellungsform möglichst flexibel gestalten zu können, wurden außerdem, wie bereits beschrieben, kein Mapping durchgeführt, das Datenpunkte zusammenfasst, sondern es wurden die expliziten Datenpunkte verwendet. Dadurch wurde außerdem die Genauigkeit der Daten aufrechterhalten. Auch hier gilt es zu beachten, dass dadurch auch Ausreißer bei kleineren Patientengrafen größere Bedeutung beigemessen werden könnte, als ihnen eigentlich zusteht. Eine Glättung beziehungsweise Zusammenfassung beispielsweise durch Mapping oder Klassifizierung der Daten könnte die Auswirkungen solcher Ausreißer abmildern, was aber wiederum zu einem Verlust der Genauigkeit der Daten führen würde. Auch hier kommt es aber auch entscheidend darauf an, wie das gewählte Ähnlichkeitsmaß die Daten bewertet und möglicherweise von sich aus die Gewichtung von Ausreißern reduziert.

Speicherverfahren

Um diese Grafen speichern zu können, wurde zunächst eine Grafdatenbank als bevorzugte Speicherform ausgewählt. Die Vorteile lagen zunächst auf der Hand: einen Grafen kann man wahrscheinlich am effizientesten in einer Grafdatenbank darstellen, die Knoten und Kanten eben auch als solche speichert. Diese

Grafdatenbanken sind eben auch für eine solche Form der Speicherung optimiert und bieten optimierte Anfragesprachen, mit deren Hilfe die Implementierung der Vergleichsalgorithmen einfacher sein sollte, als komplexe Join-Abfragen einer relationalen SQL-Datenbank. Aus diesem Grund wurde zunächst eine Grafdatenbank unter einigen möglichen Kandidaten ausgewählt, die für diese Arbeit die optimale Wahl darstellt. Die Wahl fiel letzten Endes auf die Grafdatenbank neo4j, da neo4j, anders als alle anderen Datenbankkandidaten im grafenbasierten Modell, die Anforderungen an ungerichtete Grafen, DML, DDL, API und GUI zur Anzeige der Daten erfüllte. Ein Geschwindigkeitsvergleich von neo4j mit dem relationalen DBMS PostgreSQL kam aber letztlich auf keinen nennenswerten Effizienz- oder Geschwindigkeitsvorteil der Grafdatenbank gegenüber der relationalen Datenbank. Dies kann unterschiedliche Gründe haben. So könnte beispielsweise der Aufbau der Grafen an sich den Effizienzvorteil der Grafdatenbanken bei größeren Grafen gegenüber relationalen Datenbanken zunichtemachen. Der Effizienzvorteil von Grafdatenbanken wird nämlich meist dann vorgebracht, wenn ein einzelner riesiger Graf betrachtet wird, bei dem die Knoten meist auch noch hochvernetzt sind. Die Grafen in dieser Arbeit sind weder alle zu einem Grafen aggregiert, sondern bilden jeder für sich einen einzelnen Grafen, noch sind die Grafenknoten hochvernetzt. So kann ein Knoten in den hier behandelten Grafen maximal vier Kanten aufweisen. Eine weitere Möglichkeit ist in der Natur des Ähnlichkeitsmaßes NetSimile begründet. Dieser Algorithmus orientiert sich an der Topologie des Gesamtgrafen, dabei geht es nicht darum einzelne Knoten in einem riesigen Grafen zu finden, wie es oft in Artikeln gemacht wird, die den Effizienzvorteil von Grafdatenbanken gegenüber relationalen Datenbanken herausstellen, sondern eben darum die Topologie des Gesamtgrafen zu analysieren, um durch die daraus berechneten Werte mehrere Grafen miteinander vergleichen zu können. Ebenso könnte aber auch ein anderes Ähnlichkeitsmaß besser in den Kontext der Grafdatenbank passen und müsste in diesem Zusammenhang dann noch einmal untersucht werden. Möglicherweise hätte man auch den NetSimile-Ansatz noch mit anderen Methoden in der Grafdatenbank untersuchen können. So hätten möglicherweise gespeicherte Prozeduren (engl. Stored Procedures) die Berechnung der Scores innerhalb der Datenbank beschleunigen können, sodass die Notwendigkeit entfallen wäre,

jeden einzelnen Knoten für die Berechnung aufzusuchen. Andererseits wurde mit dem Aufbau der relationalen Datenbank, die an die Struktur der Grafdatenbank angelehnt wurde, ein System geschaffen, das eine annehmbare Geschwindigkeit bei der Berechnung liefert, weshalb auch dieses Speicherverfahren seine Daseinsberechtigung hat. Möglicherweise kann aber durch eine Optimierung der Grafdatenbank, wie oben genannt, ein weiterer Geschwindigkeitsvorteil erzielt werden. Diese Möglichkeit kann in zukünftigen Arbeiten untersucht werden und dient einer weiteren Optimierung des gesamten Entscheidungsunterstützungssystems.

Auswahl und Anwendung eines Ähnlichkeitsmaßes

Um die zuvor erstellten Grafen miteinander vergleichen zu können und bestimmen zu können, welche Grafen und damit welche Patienten einander ähnlich sind, wurde die Auswahl eines Algorithmus beschrieben, der in der Grafentheorie verankert ist und aufgrund seiner Eigenschaften für das hier dargestellte Problem des Vergleichs zweier Patientengrafen gut geeignet erscheint. Dieses Auswahlvorgehen wurde in Kapitel 4.4 beschrieben. Dabei kam eine unsystematische Literaturrecherche zum Einsatz, die auf den Ergebnissen der in Kapitel 4.1 beschriebenen systematischen Literaturrecherche basiert. Dieser Umstand bedeutet natürlich gleichzeitig, dass für die Literaturrecherche aus Kapitel 4.4 dieselben Einschränkungen gelten wie für die Literaturrecherche aus Kapitel 4.1. Bei der unsystematischen Literaturrecherche wurden, wie in Kapitel 3.4 beschrieben, die Quellen der Artikel aus der systematischen Literaturrecherche genutzt, um weitere Artikel zu Grafalgorithmen zu identifizieren. Dabei wurden rekursiv die Quellen dieser Artikel nach weiteren Artikeln durchsucht, die Grafalgorithmen beschreiben oder auf solche hinweisen. Dieses Verfahren ist direkt durch die Auswahl der Quellen eingeschränkt, die die Autoren in ihren Artikeln nutzen. Dadurch kommen Artikel nicht zum Tragen, die nicht zitiert wurden und ebenso entfallen dadurch auch alle Artikel, die neuer sind als die Ausgangsartikel. Des Weiteren muss bei diesem Vorgehen von der Prämisse ausgegangen werden, dass die Autoren der entsprechenden Artikel sorgfältig recherchiert haben. Alles in allem bietet dieses Verfahren gewisse Einschränkungen gegenüber einer

systematischen Literaturrecherche. Nichtsdestotrotz wurden so einige mögliche Kandidaten-Grafvergleichsalgorithmen identifiziert, sodass der Vergleich der Algorithmen durchaus aussagekräftig ist. Bei der Betrachtung dieser Ähnlichkeitsalgorithmen blieb am Ende lediglich ein Kandidat übrig, der die gesetzten Anforderungen an ein Ähnlichkeitsmaß erfüllt hat. Deshalb stellt sich an dieser Stelle die Frage, ob die Anforderungen an ein Ähnlichkeitsmaß vielleicht zu eng gefasst wurden und ob ein im Kapitel 4.4 ausgeschlossener Algorithmus nicht vielleicht doch passende Ergebnisse geliefert hätte, wenn man auch die Darstellungsform des Grafen anders gewählt hätte.

5.3 Ausblick

In Kapitel 5.1 konnte bereits gezeigt werden, dass die vorliegende Arbeit die Grundlage für die Entwicklung eines Entscheidungsunterstützungssystems bilden kann, welches Mediziner bei Entscheidungen hinsichtlich Diagnose und Therapie unterstützen kann. Grundlage für die Arbeit war eine Literaturstudie, in der unterschiedliche Artikel mit Patientengrafen untersucht wurden. Die unterschiedlichen Felder (Hirntumore, Nierenversagen, Patienten im Generellen und so weiter), die in den Artikeln in der Literaturstudie untersucht wurden, zeigen, dass viel Potenzial für weitere Studien in diesem Forschungsfeld existiert. Die Möglichkeiten eines Systems, das Grafen zur Darstellung individueller Patientendaten nutzt, sind sehr breit gefächert und eröffnen neue Chancen, beispielsweise im klinischen Kontext. Diese Arbeit beschäftigt sich unter anderem mit der Möglichkeit der Nutzung der Grafen als Entscheidungsunterstützungssystem. Um dieses Ziel erreichen zu können, sind im Anschluss an diese Arbeit verschiedene Möglichkeiten zur Verbesserung und Weiterentwicklung des Systems denkbar. So könnte durch Entwicklung eines sehr spezifischen Ähnlichkeitsmaßes die Identifizierung bestimmter Krankheitsbilder verbessert werden. Das kann bei relativ unspezifischen Krankheitsbildern durchaus von Vorteil sein, auch wenn die Spezialisierung auf ein bestimmtes Krankheitsbild im Algorithmus andere Krankheitsbilder zunächst außen vorlässt. Zu solchen relativ unspezifischen und schwer zu identifizierenden Krankheitsbildern gehört die Sepsis. In dieser Arbeit konnte bereits gezeigt werden, dass mit dem System Sepsis-Patienten zu einem

gewissen Grad identifiziert werden können. Darauf aufbauend könnte man verschiedene Optimierungen testen, die eine verbesserte Identifizierung möglich machen könnten. Um eine solche Verbesserung der Genauigkeit der Ähnlichkeitsberechnung erreichen zu können, können an unterschiedlichen Stellen Veränderungen vorgenommen werden. So kann aus der jetzigen Gesamtgrafenanalyse eine Teilgrafenanalyse gemacht werden. Diese Teilgrafen werden dann bei verschiedenen Patienten untereinander verglichen. Diese Möglichkeit würde sich anbieten, wenn die Untersuchung der Ähnlichkeiten in eine Richtung führen würde, in der man nur auf bestimmte Krankheitsbilder hin untersuchen will. Des Weiteren können die gesammelten Daten eines Patienten getrennt nach Krankenhausaufenthalten betrachtet werden oder aber die Daten eines Patienten werden komplett zusammen betrachtet werden. Dabei muss zum Teil der Einzelfall betrachtet werden. So ist es möglich, dass für einen Patienten beispielsweise zwei Krankenhausaufenthalte gespeichert wurden, die voneinander unabhängig sind, da sie zum Beispiel zwei völlig unterschiedliche und voneinander unabhängige Erkrankungen des Patienten behandeln. Bei einem zweiten Patienten könnte dagegen die Möglichkeit bestehen, dass er beispielsweise zwei oder mehr Krankenhausaufenthalte hat, die aber alle aufeinander aufbauen und voneinander abhängig sind. So könnte die erste Krankheit die zweite bedingen und die ersten beiden zusammengenommen die dritte und so weiter. Hierbei gilt es zu untersuchen, inwieweit die Trennung dieser Daten einen Informationsverlust oder eine Verbesserung der Ergebnisse durch die getrennte Betrachtung von möglicherweise unabhängigen Daten darstellt. Eine weitere Stellschraube für das vorgestellte System stellt die Einbeziehung der Chartevents dar. So gibt es Krankheitsbilder, die in Bezug beispielsweise auf den Herzschlag definiert werden. Eine Sepsis wird so beispielsweise teilweise über einen erhöhten Herzschlag definiert. Die Erfassung der Herzschläge pro Minute und anderer Vitalzeichen wird getrennt von den Laborwerten gespeichert. Ersteres wird in der Tabelle chartevents gespeichert, letzteres in der Tabelle labevents. Eine Einbeziehung der Chartevents kann zu einem Informationsgewinn führen, kann aber gleichzeitig die Ergebnisse verwässern, dadurch, dass normale Chartevents überproportional repräsentiert werden gegenüber anormalen Chartevents. Durch die häufige Messung der Vitalzeichen und die zeitnahe Einleitung von Gegenmaßnahmen bei schlechten

Vitalzeichen durch die Klinikbelegschaft, kann es vorkommen, dass auf einige tausend Messwerte nur eine Hand voll anormaler Messwerte derselben Messwertkategorie kommen, wenn überhaupt. Mit dem aktuell gewählten Algorithmus bedeutet dies, dass die Struktur der Chartevents bei Nutzung von NetSimile in den meisten Fällen sehr ähnliche Werte bei der Strukturanalyse der Chartevents im Grafen liefern kann. Ob die Chartevents dann überhaupt noch ausschlaggebende Verbesserungen bei der Identifizierung von ähnlichen Patienten beitragen können, gilt es zu untersuchen. Möglicherweise ist eine Abwandlung oder ein Austausch des aktuellen Ähnlichkeitsmaßes notwendig, um die Chartevents gewinnbringend in die Untersuchung mit einfließen zu lassen. So könnte eine mögliche Abwandlung des NetSimile-Algorithmus dahingehend Verbesserungen zur Folge haben, dass die Gewichtung der anormalen Messwerte angepasst werden kann. Im Anschluss an die Optimierung des Ähnlichkeitsmaßes, sodass ausreichend genaue Ergebnisse für ein Entscheidungsunterstützungssystem generiert werden können, muss eine grafische Benutzeroberfläche entwickelt werden, die es Anwendern ermöglicht, möglichst schnell und benutzerfreundlich ähnliche Patienten, Diagnosen und/oder Therapien angezeigt zu bekommen. Eine mögliche Herangehensweise wäre die Summierung der vorkommenden Diagnosen in den ähnlichsten Patienten und eine entsprechend sortierte Auflistung der Diagnosen, sodass der Anwender aus den Diagnosen wählen kann, die bei den ähnlichsten Patienten am häufigsten vorkommen. Daneben könnten auch einfach die ähnlichsten Patienten gelistet werden, sortiert nach ihrer Ähnlichkeit. Dadurch kann eine mögliche Überrepräsentierung der häufigsten Diagnosen durch die zusätzliche Anzeige der ähnlichsten Patienten ausgeglichen werden. Im Allgemeinen kann dazu auch gesagt werden, dass das System im klinischen Alltag bei vielen Patienten wahrscheinlich keine eindeutige Lösung vorschlagen kann, sondern lediglich eine Reihe von möglichen Diagnosen beziehungsweise Patienten, die dem untersuchten Patienten ähnlich sind. Das System bietet dem Mediziner durch diese Vorschläge aber die Möglichkeit auf eventuell nicht bedachte Diagnosen oder Therapien zu testen und könnte so dazu beitragen, dass Diagnosen schneller und besser gestellt werden können oder erfolgsversprechende Therapieansätze besser erkannt werden könnten, da sie bei anderen Patienten bereits zum Erfolg geführt haben.

6 Zusammenfassung

Entscheidungsunterstützungssysteme sind rechnerbasierte Anwendungssysteme, die menschlichen Entscheidern in komplexen Situationen bei der Entscheidungsfindung helfen sollen. Diese EUS sollen im klinischen Alltag dabei helfen Früherkennung, Diagnostik und Therapieentscheidung für Patienten zu verbessern. Zur Umsetzung eines solchen Entscheidungsunterstützungssystems wurden in dieser Arbeit die Grundlagen geschaffen. Für gewöhnlich liegen die Daten von Patienten und Behandlungsverläufen in Tabellen vor. Dabei werden allerdings die temporalen Zusammenhänge der Daten außen vor gelassen. Ziel dieser Arbeit war die Erarbeitung einer Darstellungsform, in der die Patientendaten in einem Grafen dargestellt werden können, die Erarbeitung einer passenden Speicherungsform sowie die Auswahl eines Ähnlichkeitsmaßes, das ähnliche Patienten mit Hilfe der Grafdarstellung identifizieren kann. Bei Nutzung der Daten in Grafen konnte in dieser Arbeit eine Darstellungsform der Patientendaten entwickelt werden, die die temporalen Zusammenhänge zwischen den Messwerten und Daten der Behandlungsverläufe aufnimmt und als Teil eines Patientengrafen darstellt. Pro Patient wird somit ein Graf von dessen Messwerten erstellt. Dabei wird auch zwischen den Krankenhausaufenthalten eines Patienten unterschieden. Im nächsten Schritt konnte ein Speicherverfahren etabliert werden, das für die weiteren Schritte auf dem Weg zum Entscheidungsunterstützungssystem die nötige Flexibilität bietet. So wurden verschiedene DBMS miteinander verglichen, wobei die Wahl des Speicherverfahren letzten Endes auf das relationale Datenbankmodell und DBMS PostgreSQL fiel. Zunächst wurde allerdings die Grafdatenbank neo4j eingesetzt, es zeigte sich aber, dass für die hier vorgestellte Anwendung kein nennenswerter Geschwindigkeitsvorteil einer grafbasierten Datenbank gegenüber einer relationalen Datenbank festgestellt werden konnte. Somit wurden die Daten im relationalen Datenbankmodell gespeichert, sodass allerdings auch die zeitlichen Zusammenhänge und Verbindungen der Knoten des Grafen mit abgespeichert werden konnten. Das Speicherverfahren wurde außerdem so gewählt, dass die originalen Messwerte sowie die originalen Abstände zwischen den Knoten erhalten blieben, auch wenn sie für das vorgestellte System nicht relevant waren. Die erzeugten

Patientengrafen wurden durch Nutzung eines in der Arbeit aus verschiedenen Algorithmen ausgewählten Algorithmus *(NetSimile)* verglichen. Durch dieses System konnte bereits ein erster Eindruck vermittelt werden, welche Möglichkeiten ein darauf aufbauendes EUS bietet, um ähnliche Patienten zu finden und daraus eine Diagnosen- oder Therapieentscheidung abzuleiten. Das Ähnlichkeitsmaß wurde auch auf seine Funktionsweise hin getestet, wobei eine Verteilung der ähnlichen Patienten untersucht wurde sowie eine Häufigkeitsanalyse der häufigsten Diagnosen der ähnlichsten Patienten durchgeführt wurde, mit dem Ziel die Grafdarstellung sowie die Anwendung des Ähnlichkeitsmaßes auf diese Grafdarstellung auf Funktionalität zu überprüfen. Die Ergebnisse legen in diesem Zusammenhang nahe, dass das Ähnlichkeitsmaß im Zusammenhang mit der Grafdarstellung funktioniert. Allerdings gibt es noch Verbesserungsbedarf bei der Umsetzung des Ähnlichkeitsmaßes und auch für die Untersuchung auf Funktionalität des Ähnlichkeitsmaßes zusammen mit der Darstellungsform sollten weitergehende Untersuchungen durchgeführt werden.

7 Literaturverzeichnis

Aguilera-Mendoza, Longendri; Marrero-Ponce, Yovani; García-Jacas, César R.; Chavez, Edgar; Beltran, Jesus A.; Guillen-Ramirez, Hugo A.; Brizuela, Carlos A. (2020): Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides. An unsupervised learning approach. In: *Scientific reports* 10 (1), S. 18074. DOI: 10.1038/s41598-020-75029-1.

Alvarez, Marco A.; Qi, Xiaojun; Yan, Changhui (2011): A shortest-path graph kernel for estimating gene product semantic similarity. In: *Journal of biomedical semantics* 2, S. 3. DOI: 10.1186/2041-1480-2-3.

Anderson, Marti J.; Ellingsen, Kari E.; McArdle, Brian H. (2006): Multivariate dispersion as a measure of beta diversity. In: *Ecology letters* 9 (6), S. 683–693. DOI: 10.1111/j.1461-0248.2006.00926.x.

Angles, Renzo (2012): A Comparison of Current Graph Database Models. In: *2012 IEEE International Conference on Data Engineering Workshops (ICDEW). DOI:* 10.1109/ICDEW.2012.31.

Angles, Renzo; Barceló, Pablo; Ríos, Gonzalo (2013): A practical query language for graph dbs. In: *Bravo, L. and Lenzerini, M. (eds.), Proceedings 7th Alberto Mendelzon International Workshop on Foundations of Data Management, CEUR Workshop Proceedings* (1087).

Assmann, H.; Bergmann, G. v.; Doerr, R.; Grafe, E.; Heilmeyer, L.; Hiller, F. et al. (1949): Lehrbuch der Inneren Medizin. Berlin, Heidelberg: Springer Berlin Heidelberg.

Atif, J.; Hudelot, C.; Nempont, O.; Richard, N.; Batrancourt, B.; Angelini, E.; Bloch, I.: GRAFIP: A FRAMEWORK FOR THE REPRESENTATION OF HEALTHY AND PATHOLOGICAL CEREBRAL INFORMATION. In: 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Arlington, VA, USA, S. 205–208.

Atif, J.; Hudelot, C.; Nempont, O.; Richard, N.; Batrancourt, B.; Angelini, E.; Bloch, I. (2007): GRAFIP. A FRAMEWORK FOR THE REPRESENTATION OF HEALTHY AND PATHOLOGICAL CEREBRAL INFORMATION. In: 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, S. 205–208.

Bean, Daniel M.; Wu, Honghan; Iqbal, Ehtesham; Dzahini, Olubanke; Ibrahim, Zina M.; Broadbent, Matthew et al. (2017): Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. In: *Scientific reports* 7 (1), S. 16416. DOI: 10.1038/s41598-017-16674-x.

Berlingerio, Michele; Koutra, Danai; Eliassi-Rad, Tina; Falousos, Christos (2012): NetSimile: A Scalable Approach to Size-Independent Network Similarity, zuletzt geprüft am 07.08.2017.

Bollobás, Bella (2010): Modern Graph Theory, 5. Print. 184. Aufl. New York, NY: Springer.

Bray, Tim (2017): The JavaScript Object Notation (JSON) Data Interchange Format: RFC Editor (Request for Comments) (8259). Online verfügbar unter https://rfc-editor.org/rfc/rfc8259.txt.

Bray, Tim; Paoli, Jean; Sperberg-McQueen, C.M.; Maler, Eve; Yergeau, Francoi (2008): Extensible Markup Language (XML) 1.0 (Fifth Edition) (1.0.5). Online verfügbar unter https://www.w3.org/TR/xml/, zuletzt aktualisiert am 26.11.2008, zuletzt geprüft am 23.03.2020.

Brown, Sherry-Ann (2016): Patient Similarity: Emerging Concepts in Systems and Precision Medicine. In: *frontiers in Physiology* 7, S. 561. DOI: 10.3389/fphys.2016.00561.

Bu, Shugui; Guo, Lilu; Li, Rongyuan; Lu, Jianbo; Zhu, Xiaoshu (2019): GSE. In: Proceedings of the 2019 4th International Conference on Intelligent Information Processing. ICIIP 2019: 2019 4th International Conference on Intelligent Information Processing. China China, 16 11 2019 17 11 2019. New York, NY, USA: ACM, S. 405–409.

Bunke, H.; Allermann, G. (1983): Inexact graph matching for structural pattern recognition. In: *Pattern Recognition Letters* 1 (4), S. 245–253. DOI: 10.1016/0167-8655(83)90033-8.

Campanharo, Andriana S. L. O.; Sirer, M. Irmak; Malmgren, R. Dean; Ramos, Fernando M.; Amaral, Luís A. Nunes (2011): Duality between time series and networks. In: *PLoS ONE* 6 (8), e23378. DOI: 10.1371/journal.pone.0023378.

Campbell, W. Scott; Pedersen, Jay; McClay, James C.; Rao, Praveen; Bastola, Dhundy; Campbell, James R. (2015): An alternative database approach for management of SNOMED CT and improved patient data queries. In: *Journal of biomedical informatics* 57, S. 350–357. DOI: 10.1016/j.jbi.2015.08.016.

Chen, Ling; Li, Xue; Sheng, Quan Z.; Peng, Wen-Chih; Bennett, John; Hu, Hsiao-Yun; Huang, Nicole (2016): Mining Health Examination Records—A Graph-Based Approach. In: *IEEE Trans. Knowl. Data Eng.* 28 (9), S. 2423–2437. DOI: 10.1109/TKDE.2016.2561278.

Chen, Yuanzhe; Xu, Panpan; Ren, Liu (2017): Sequence Synopsis: Optimize Visual Summary of Temporal Event Data. In: *IEEE transactions on visualization and computer graphics. DOI:* 10.1109/TVCG.2017.2745083.

Classen, Meinhard; Diehl, Volker; Kochsiek, Kurt (Hg.) (2010): Innere Medizin. 1200 Tabellen, 200 Kasuistiken, 450 Zusammenfassungen, 180 Praxisfragen ; [mit dem Plus im Web]. 6., komplett überarb. Aufl., [Nachdr.]. München: Elsevier Urban & Fischer. Online verfügbar unter http://institut.elsevierelibrary.de/product/innere-medizin3948#.UrLIBlPgsVo.

Clermont, Gilles; Kaplan, Vladimir; Moreno, Rui; Vincent, Jean-Louis; Linde-Zwirble, Walter T.; van Hout, Ben; Angus, Derek C. (2004): Dynamic microsimulation to model multiple outcomes in cohorts of critically ill patients.

In: *Intensive care medicine* 30 (12), S. 2237–2244. DOI: 10.1007/s00134-004-2456-5.

Codd, E. F. (1970): A relational model of data for large shared data banks. In: *Commun. ACM* 13 (6), S. 377–387. DOI: 10.1145/362384.362685.

Codd, Edgar F. (1991): The relational model for database management. Version 2. Reprinted with corr. Reading, Mass.: Addison-Wesley.

Dankar, Fida Kamal; El Emam, Khaled (2013): Practicing differential privacy in health care. A review. In: *Trans. Data Priv.* 6 (1), S. 35–67.

Dominguez-Sal, D.; Urbón-Bayes, P.; Giménez-Vanó, A.; Gómez-Villamor, S.; Martínez-Bazán, N.; Larriba-Pey, J. L. (2010): Survey of Graph Database Performance on the HPC Scalable Graph Analysis Benchmark. In: *Shen H.T. et al. (eds) Web-Age Information Management. WAIM 2010. Lecture Notes in Computer Science* (6185).

Dreiher, Jacob; Almog, Yaniv; Sprung, Charles L.; Codish, Shlomi; Klein, Moti; Einav, Sharon et al. (2012): Temporal trends in patient characteristics and survival of intensive care admissions with sepsis: a multicenter analysis*. In: *Critical care medicine* 40 (3), S. 855–860. DOI: 10.1097/CCM.0b013e318236f7b8.

Edlich, Stefan; Friedland, Achim; Hampel, Jens; Brauer, Benjamin (2010): NoSQL. Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. München: Hanser. Online verfügbar unter http://sub-hh.ciando.com/book/?bok_id=47525.

Eisentraut, Peter; Helmle, Bernd (2013): PostgreSQL-Administration. [die fortschrittlichste Open-Source-Datenbank ; behandelt PostgreSQL 9.2]. 3. Aufl. Beijing: O'Reilly.

Ercan, Mehmet; Lane, Michael (2014): An Evaluation of NoSQL Databases for Electronic Health Record Systems.

Esteban, Cristóbal; Schmidt, Danilo; Krompass, Denis; Tresp, Volker (2015): Predicting Sequences of Clinical Events by Using a Personalized Temporal Latent Embedding Model. In: *2015 International Conference on Healthcare Informatics*, S. 130–139.

Fabregat, Antonio; Korninger, Florian; Viteri, Guilherme; Sidiropoulos, Konstantinos; Marin-Garcia, Pablo; Ping, Peipei et al. (2018): Reactome graph database. Efficient access to complex pathway data. In: *PLoS computational biology* 14 (1), e1005968. DOI: 10.1371/journal.pcbi.1005968.

Fleischmann, Carolin; Thomas-Rueddel, Daniel O.; Hartmann, Michael; Hartog, Christiane S.; Welte, Tobias; Heublein, Steffen et al. (2016): Hospital Incidence and Mortality Rates of Sepsis. In: *Deutsches Arzteblatt international* 113 (10), S. 159–166. DOI: 10.3238/arztebl.2016.0159.

Fleischmann-Struzek, C.; Schwarzkopf, D.; Reinhart, K. (2022): Inzidenz der Sepsis in Deutschland und weltweit. Aktueller Wissensstand und Limitationen der Erhebung in Abrechnungsdaten. In: *Medizinische Klinik, Intensivmedizin und Notfallmedizin* 117 (4), S. 264–268. DOI: 10.1007/s00063-021-00777-5.

Friedman, Carol; Hripcsak, George; Johnson, Stephen B.; Cimino, James J.; Clayton, Paul D. (1990): A Generalized Relational Schema for an Integrated Clinical Patient Database. In: *Proc Annu Symp Comput Appl Med Care*, S. 335–339.

Fries, James F. (1972): Time-Oriented Patient Records and a Computer Databank. In: *JAMA* 222 (12), S. 1536–1542. DOI: 10.1001/jama.1972.03210120034009.

Hanzlicek, Petr; Spidlen, Josef; Nagy, Miroslav (2004): Universal electronic health record MUDR. In: *Studies in health technology and informatics* 105, S. 190–201.

Herold, Gerd (Hg.) (2017): Innere Medizin 2017. Eine vorlesungsorientierte Darstellung : unter Berücksichtigung des Gegenstandskataloges für die Ärztliche Prüfung : mit ICD 10-Schlüssel im Text und Stichwortverzeichnis. Dr. Gerd Herold. Köln: Gerd Herold.

Holme, Petter (2015): Modern temporal network theory. A colloquium. In: *Eur. Phys. J. B* 88 (9), S. 558. DOI: 10.1140/epjb/e2015-60657-4.

Hu, Wei; Fu, Zeqing; Guo, Zongming (2019): Local Frequency Interpretation and Non-Local Self-Similarity on Graph for Point Cloud Inpainting. In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society. DOI:* 10.1109/TIP.2019.2906554.

Huber, Wolfgang; Carey, Vincent J.; Long, Li; Falcon, Seth; Gentleman, Robert (2007): Graphs in molecular biology. In: *BMC bioinformatics* 8 Suppl 6, S. S8. DOI: 10.1186/1471-2105-8-S6-S8.

Idziaszek, Przemysław; Mueller, Wojciech; Rudowicz-Nawrocka, Janina; Gruszczyński, Michał; Kujawa, Sebastian; Górna, Karolina; Balcerzak, Kinga (2016): Visualisation of Relational Database Structure by Graph Database. In: *CMST* 22 (4), S. 217–224. DOI: 10.12921/cmst.2016.0000014.

ISO ISO/IEC 9075 -1:2016, 2016: Information technology - Database languages - SQL - Part 1: Framework (SQL/Framework).

Jaiswal, Garima; Agrawal, Arun Prakash (2015): Comparative analysis of Relational and Graph databases. In: *IJIACS - International Journal of Innovations & Advancement in Computer Science* 4, S. 181–183. Online verfügbar unter http://www.academicscience.co.in/admin/resources/project/paper/f2015022514 24869936.pdf, zuletzt geprüft am 14.08.2018.

Jiang, Chuntao; Coenen, Frans; Zito, Michele (2013): A survey of frequent subgraph mining algorithms. In: *The Knowledge Engineering Review* 28 (1), S. 75–105. DOI: 10.1017/S0269888912000331.

Johnpaul, C. I.; Mathew, Tojo: A Cypher query based NoSQL data mining on protein datasets using Neo4j graph database. In: 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), S. 1–6.

Johnson, Alistair E. W.; Pollard, Tom J.; Shen, Lu; Lehman, Li-Wei H.; Feng, Mengling; Ghassemi, Mohammad et al. (2016): MIMIC-III, a freely accessible critical care database. In: *Scientific data* 3, S. 160035. DOI: 10.1038/sdata.2016.35.

Kaur, Karamjit; Rani, Rinkle (2015): Managing Data in Healthcare Information Systems. Many Models, One Solution. In: *Computer* 48 (3), S. 52–59. DOI: 10.1109/MC.2015.77.

Klein, John; Gorton, Ian; Ernst, Neil; Donohoe, Patrick; Pham, Kim; Matser, Chrisjan (2015): Performance Evaluation of NoSQL Databases. In: Rekha Singhal und Dheeraj Chahal (Hg.): Proceedings of the 1st Workshop on Performance Analysis of Big Data Systems - PABS '15. the 1st Workshop. Austin, Texas, USA, 01.02.2015 - 01.02.2015. New York, New York, USA: ACM Press, S. 5–10.

Koutra, Danai; Parikh, Ankur; Ramdas, Aaditya; Xiang, Jing (2011): Algorithms for graph similarity and subgraph matching. In: Proc. Ecol. Inference Conf, Bd. 17.

Kuckartz, Udo (2014): Qualitative Text Analysis. A Guide to Methods, Practice & Using Software. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd.

Lacasa, Lucas; Luque, Bartolo; Ballesteros, Fernando; Luque, Jordi; Nuño, Juan Carlos (2008): From time series to complex networks. The visibility graph. In: *Proceedings of the National Academy of Sciences of the United States of America* 105 (13), S. 4972–4975. DOI: 10.1073/pnas.0709247105.

Levy, Mitchell M.; Fink, Mitchell P.; Marshall, John C.; Abraham, Edward; Angus, Derek; Cook, Deborah et al. (2003): 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. In: *Critical care medicine* 31 (4), S. 1250–1256. DOI: 10.1097/01.CCM.0000050454.01978.3B.

Lin, Zijie; Gao, Liangliang; Hu, Xuexian; Zhang, Yuxuan; Liu, Wenfen (2019): Differentially Private Graph Clustering Algorithm Based on Structure Similarity. In: Proceedings of the 2019 the 9th International Conference on Communication and Network Security. New York, NY, USA: Association for Computing Machinery (ICCNS 2019), S. 63–68. Online verfügbar unter https://doi.org/10.1145/3371676.3371693.

Liu, Chuanren; Wang, Fei; Hu, Jianying; Xiong, Hui (2015): Temporal Phenotyping from Longitudinal Electronic Health Records. A Graph Based Framework. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery (KDD '15), S. 705–714. Online verfügbar unter https://doi.org/10.1145/2783258.2783352.

López, Daniela N.; Camus, Patricio A.; Valdivia, Nelson; Estay, Sergio A. (2019): Integrating species and interactions into similarity metrics. A graph theory-based approach to understanding community similarity. In: *PeerJ* 7, e7013. DOI: 10.7717/peerj.7013.

Luo, Ligang; Li, Liping; Hu, Jiajia; Wang, Xiaozhe; Hou, Boulin; Zhang, Tianze; Zhao, Lue Ping (2016): A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. In: *BMC medical informatics and decision making* 16, S. 114. DOI: 10.1186/s12911-016-0357-5.

Luque, B.; Lacasa, L.; Ballesteros, F.; Luque, J. (2009): Horizontal visibility graphs. Exact results for random time series. In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 80 (4 Pt 2), S. 46103. DOI: 10.1103/PhysRevE.80.046103.

Lysenko, Artem; Roznovăţ, Irina A.; Saqi, Mansoor; Mazein, Alexander; Rawlings, Christopher J.; Auffray, Charles (2016): Representing and querying disease networks using graph databases. In: *BioData mining* 9, S. 23. DOI: 10.1186/s13040-016-0102-8.

Miller, Justin J. (2013): Graph database applications and concepts with Neo4j. In: Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA, Bd. 2324.

Moher, David; Liberati, Alessandro; Tetzlaff, Jennifer; Altman, Douglas G. (2009): Preferred reporting items for systematic reviews and meta-analyses. The PRISMA statement. In: *BMJ (Clinical research ed.)* 339, b2535. DOI: 10.1136/bmj.b2535.

Moniruzzaman, A. B. M.; Hossain, Syed Akhter (2013): NoSQL Database. New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. In: *CoRR* abs/1307.0191.

Moskovitch, Robert; Shahar, Yuval (2009): Medical temporal-knowledge discovery via temporal abstraction. In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2009, S. 452–456.

Müller, R.; Thews, O.; Rohrbach, C.; Sergl, M.; Pommerening, K. (1996): A graph-grammar approach to represent causal, temporal and other contexts in an oncological patient record. In: *Methods of information in medicine* 35 (2), S. 127–141.

Nayak, A.; Poriya, A.; Poojary, Dikshay (2013): Article. Type of nosql databases and its comparison with relational databases. In: *International Journal of Applied Information Systems* 5, S. 16–19.

neo4j (Hg.) (2017): Cypher. Online verfügbar unter http://neo4j.com/docs/developer-manual/current/cypher/#cypher-intro, zuletzt geprüft am 17.10.2017.

Newman, M. E. J. (2016): Networks. An introduction. Reprinted. Oxford: Oxford University Press.

Onoue, Y.; Kukimoto, N.; Sakamoto, N.; Koyamada, K. (2016): Minimizing the Number of Edges via Edge Concentration in Dense Layered Graphs. In: *IEEE transactions on visualization and computer graphics* 22 (6), S. 1652–1661. DOI: 10.1109/TVCG.2016.2534519.

Pavlopoulos, Georgios A.; Secrier, Maria; Moschopoulos, Charalampos N.; Soldatos, Theodoros G.; Kossida, Sophia; Aerts, Jan et al. (2011): Using graph theory to analyze biological networks. In: *BioData mining* 4, S. 10. DOI: 10.1186/1756-0381-4-10.

Puentes, John; Batrancourt, Bénédicte; Atif, Jamal; Angelini, Elsa; Lecornu, Laurent; Zemirline, Abdelhamid et al. (2008): Integrated multimedia electronic patient record and graph-based image information for cerebral tumors. In: *Computers in biology and medicine* 38 (4), S. 425–437. DOI: 10.1016/j.compbiomed.2008.01.009.

Rascovsky, Simón J.; Delgado, Jorge A.; Sanz, Alexander; Calvo, Víctor D.; Castrillón, Gabriel (2012): Informatics in radiology. Use of CouchDB for document-based storage of DICOM objects. In: *Radiographics : a review publication of the Radiological Society of North America, Inc* 32 (3), S. 913–927. DOI: 10.1148/rg.323115049.

Raymond, John W.; Willett, Peter (2002): Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. In: *Journal of computer-aided molecular design* 16 (1), S. 59–71. DOI: 10.1023/a:1016387816342.

Rieß, Werner; Wirtz, Markus Antonius; Barzel, Bärbel; Schulz, Andreas; Altenburger, Pia (Hg.) (2012): Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. Schüler lernen wissenschaftlich denken und arbeiten. Münster: Waxmann. Online verfügbar unter http://www.content-select.com/index.php?id=bib_view&ean=9783830976875.

Riggs, Simon (2015): PostgreSQL 9 administration cookbook. Over 150 recipes to help you run an efficient PostgreSQL database in the cloud. Second edition. Birmingham, UK: Packt Publishing. Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=987632.

Saeed, M.; Lieu, C.; Raber, G.; Mark, R. G. (2002): MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: *Computers in cardiology* 29, S. 641–644.

Sankaran, R. T.; Mattana, J.; Pollack, S.; Bhat, P.; Ahuja, T.; Patel, A.; Singhal, P. C. (1997): Laboratory abnormalities in patients with bacterial pneumonia. In: *Chest* 111 (3), S. 595–600. DOI: 10.1378/chest.111.3.595.

Schmoch, T.; Bernhard, M.; Uhle, F.; Gründling, M.; Brenner, T.; Weigand, M. A. (2017): Neue SEPSIS-3-Definition. Müssen wir Sepsis in Zukunft behandeln, bevor wir sie diagnostizieren dürfen? In: *Der Anaesthesist. DOI:* 10.1007/s00101-017-0316-2.

Schrodt, Jens; Dudchenko, Aleksei; Knaup-Gregori, Petra; Ganzinger, Matthias (2020): Graph-Representation of Patient Data. A Systematic Literature Review. In: *Journal of medical systems* 44 (4), S. 86. DOI: 10.1007/s10916-020-1538-4.

Sharma, Harshita; Alekseychuk, Alexander; Leskovsky, Peter; Hellwich, Olaf; Anand, R. S.; Zerbe, Norman; Hufnagl, Peter (2012): Determining similarity in

histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. In: *Diagnostic pathology* 7, S. 134. DOI: 10.1186/1746-1596-7-134.

Shtar, Guy; Rokach, Lior; Shapira, Bracha (2019): Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures. In: *PLoS ONE* 14 (8), e0219796. DOI: 10.1371/journal.pone.0219796.

Siek, Jeremy; Lumsdaine, Andrew; Lee, Lie-Quan (2002): The Boost graph library. User guide and reference manual. Boston: Addison-Wesley (C++ in-depth series). Online verfügbar unter http://proquest.tech.safaribooksonline.de/9780321601629.

Soulakis, Nicholas D.; Carson, Matthew B.; Lee, Young Ji; Schneider, Daniel H.; Skeehan, Connor T.; Scholtens, Denise M. (2015): Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. In: *Journal of the American Medical Informatics Association : JAMIA* 22 (2), S. 299–311. DOI: 10.1093/jamia/ocu017.

Stearns-Kurosawa, Deborah J.; Osuchowski, Marcin F.; Valentine, Catherine; Kurosawa, Shinichiro; Remick, Daniel G. (2011): The pathogenesis of sepsis. In: *Annual review of pathology* 6, S. 19–48. DOI: 10.1146/annurev-pathol-011110-130327.

UCLA (Hg.) (2022): WHAT'S WITH THE DIFFERENT FORMULAS FOR KURTOSIS? Online verfügbar unter https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-whats-with-the-different-formulas-for-kurtosis/, zuletzt geprüft am 06.01.2023.

Ugander, Johan; Karrer, Brian; Backstrom, Lars; Marlow, Cameron (2011): The anatomy of the facebook social graph. In: *arXiv preprint arXiv:1111.4503*.

Ullmann, J. R. (1976): An Algorithm for Subgraph Isomorphism. In: *J. ACM* 23 (1), S. 31–42. DOI: 10.1145/321921.321925.

VERBI Software GmbH (2018): MAXQDA. Software for qualitative and mixed methods research. Version : VERBI Software GmbH.

Vicknair, Chad; Macias, Michael; Zhao, Zhendong; Nan, Xiaofei; Chen, Yixin; Wilkins, Dawn (2010): A Comparison of a Graph Database and a Relational Database. A Data Provenance Perspective. In: H. Conrad Cunningham (Hg.): Proceedings of the 48th Annual Southeast Regional Conference. New York, NY: ACM. Online verfügbar unter http://delivery.acm.org/10.1145/1910000/1900067/a42-vicknair.pdf?ip=129.206.126.110&id=1900067&acc=ACTIVE%20SERVICE&key=2BA2C432AB83DA15%2E4992EA3396EC4E12%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&__acm__=1534235597_7cd6eba604f020e9908e08e24f2cb514, zuletzt geprüft am 14.08.2018.

Werdan, Karl; Müller-Werdan, Ursula; Schuster, Hans-Peter; Brunkhorst, Frank M. (2016): Sepsis und MODS. Berlin, Heidelberg: Springer Berlin Heidelberg.

Werners, Brigitte (Hg.) (2008): Grundlagen des Operations Research. Berlin, Heidelberg: Springer Berlin Heidelberg (Springer-Lehrbuch).

Wong, Hector R.; Weiss, Scott L.; Giuliano, John S.; Wainwright, Mark S.; Cvijanovich, Natalie Z.; Thomas, Neal J. et al. (2014): The Temporal Version of the Pediatric Sepsis Biomarker Risk Model. In: *PLoS ONE* 9 (3). DOI: 10.1371/journal.pone.0092121.

Xu, Wei; Zhou, Zhonghua; Zhou, Hong; Zhang, Wu; Xie, Jiang (2014): MongoDB Improves Big Data Analysis Performance on Electric Health Record System. In: Shiwei Ma, Li Jia, Xin Li, Ling Wang, Huiyu Zhou und Xin Sun (Hg.): Life System Modeling and Simulation, Bd. 461. Berlin, Heidelberg: Springer Berlin Heidelberg (Communications in Computer and Information Science), S. 350–357.

Xu, Xiaowei; Yuruk, Nurcan; Feng, Zhidan; Schweiger, Thomas A. J. (2007): SCAN. In: Pavel Berkhin, Rich Caruana und Xindong Wu (Hg.): Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07. the 13th ACM SIGKDD international conference. San Jose, California, USA, 12.08.2007 - 15.08.2007. New York, New York, USA: ACM Press, S. 824.

Yang, Qi; Ma, Zhan; Xu, Yiling; Li, Zhu; Sun, Jun (2020): Inferring Point Cloud Quality via Graph Similarity. In: *IEEE transactions on pattern analysis and machine intelligence* PP. DOI: 10.1109/TPAMI.2020.3047083.

Yoon, Byoung-Ha; Kim, Seon-Kyu; Kim, Seon-Young (2017): Use of Graph Database for the Integration of Heterogeneous Biological Data. In: *Genomics & informatics* 15 (1), S. 19–27. DOI: 10.5808/GI.2017.15.1.19.

Yousefi, A.; Mastouri, N.; Sartipi, K. (2009): Scenario-oriented information extraction from electronic health records. In: 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, S. 1–5.

Zador, Zsolt; Landry, Alexander; Cusimano, Michael D.; Geifman, Nophar (2019): Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care. In: *Critical care (London, England)* 23 (1), S. 247. DOI: 10.1186/s13054-019-2486-6.

Zhang, Jinghe; Gong, Jiaqi; Barnes, Laura: HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). Philadelphia, PA, USA, S. 214–221.

Zhang, Jinghe; Gong, Jiaqi; Barnes, Laura (2017): HCNN. Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). Philadelphia, PA, USA, 17.07.2017 - 19.07.2017: IEEE, S. 214–221.

Zhang, S.; Liu, L.; Li, H.; Xiao, Z.; Cui, L. (2016): MTPGraph. A Data-Driven Approach to Predict Medical Risk Based on Temporal Profile Graph. In: 2016 IEEE Trustcom/BigDataSE/ISPA, S. 1174–1181.

Zhang, Shuai; Liu, Lei; Li, Hui; Xiao, Zongshui; Cui, Lizhen: MTPGraph: A Data-Driven Approach to Predict Medical Risk Based on Temporal Profile Graph. In: 2016 IEEE Trustcom/BigDataSE/ISPA. Tianjin, China, S. 1174–1181.

Zhang, Zelin; Xu, Jinyu; Zhou, Xiao (2019): Mapping time series into complex networks based on equal probability division. In: *AIP Advances* 9 (1), S. 15017. DOI: 10.1063/1.5062590.

Zhou, Xiaohua; Han, Hyoil; Chankai, Isaac; Prestrud, Ann; Brooks, Ari (2006): Approaches to text mining for clinical medical records. In: Hisham M. Haddad (Hg.): Proceedings of the 2006 ACM symposium on Applied computing - SAC '06. the 2006 ACM symposium. Dijon, France, 23.04.2006 - 27.04.2006. New York, New York, USA: ACM Press, S. 235.

8 Eigenanteil an Datenerhebung und –Auswertung und eigene Veröffentlichungen

Die Entwicklung der Methode und der Programme sowie deren Auswertung wurden von mir durchgeführt. Die entwickelte Methode zur Darstellung von Patienten als Grafen ist das zentrale Ergebnis dieser Dissertation.

Grundlage dieser Arbeit war ein vorab publizierter Aufsatz:

Schrodt, Jens; Dudchenko, Aleksei; Knaup-Gregori, Petra; Ganzinger, Matthias (2020): Graph-Representation of Patient Data. A Systematic Literature Review. In: *Journal of medical systems* 44 (4), S. 86. DOI: 10.1007/s10916-020-1538-4.

Die oben genannte Publikation ist Grundlage dieser Arbeit. Basierend auf einer in der Publikation durchgeführten systematischen Literaturstudie wurden Publikationen identifiziert, die ein ähnliches Themenfeld bearbeiten, um die Dissertation in den entsprechenden wissenschaftlichen Kontext einzubetten. Mein Eigenanteil an der Publikation erstreckt sich auf die Erstellung der Queries zur systematischen Literaturstudie, sowie die Tätigkeiten als einer der Gutachter sowie als Gutachter im zweiten Schritt. Ebenso wurde von mir die Volltextanalyse der in die Analyse eingeschlossenen Artikel durchgeführt, sowie das Schreiben des Manuskriptentwurfs. Herr Ganzinger und Herr Dudchenko treten in dieser Publikation ebenso als Gutachter auf.

Weitere eigene Veröffentlichungen

- Ganzinger M, **Schrodt J**, Knaup P. A Concept for Graph-Based Temporal Similarity of Patient Data. Stud Health Technol Inform. 2019 Aug 21;264:138-142. doi: 10.3233/SHTI190199. PMID: 31437901.
- Pertl-Obermeyer, H., Wu, X., **Schrodt, J.**, Müdsam, C., Obermeyer, G., Schulze, W. X. (2016). Identification of Cargo for Adaptor Protein (AP) Complexes 3 and 4 by Sucrose Gradient Profiling**.** In: *Molecular & cellular proteomics: MCP* 15(9), S.2877-2889. DOI: 10.1074/mcp.M116.060129

9 Anhang

Lebenslauf

Personalien

Name:	Jens Schrodt
Geburtsdatum:	21.02.1992
Geburtsort:	Mühlacker
Staatsangehörigkeit:	deutsch
Familienstand:	verheiratet, drei Kinder

Schulischer Werdegang

1998-2002	Grundschule Schützingen
2002-2008	Salzach Gymnasium Maulbronn
2008-2011	Theodor-Heuss-Gymnasium Mühlacker
27.05.2011	Allgemeine Hochschulreife

Universitärer Werdegang

WS 2011/2012	Beginn Bachelor Biologie B. Sc. an der Universität Hohenheim
SS 2012	Parallel Beginn Bachelor Wirtschaftswissenschaften B.Sc. an der Fernuniversität Hagen (Fernstudium)
04.08.2014	Abschluss Bachelor Biologie B. Sc.
WS 2014/2015	Beginn Master Molucular Biosciences im Major Systemsbiology an der Universität Heidelberg
03.02.2017	Abschluss Master Molecular Biosciences, Major Systemsbiology

02.06.2020	Abschluss Bachelor Wirtschaftswissenschaften B. Sc. Fernuniversität Hagen

Beruflicher Werdegang

Seit 2006	(Abteilungsleiter) Softwareentwicklung Schrodt Informatik GmbH Illingen-Schützingen
Seit 08.01.2020	Geschäftsführer Schrodt Informatik GmbH
2017-2021	Hilfswissenschaftler Institut für Medizinische Informatik Universitätsklinikum Heidelberg
2017-2023	Doktorand Institut für Medizinische Informatik Universitätsklinikum Heidelberg

Danksagung

Frau Professorin Dr. Petra Knaup-Gregori und Herrn Dr. Matthias Ganzinger danke ich für die Überlassung des Themas dieser Arbeit, die Möglichkeit diese Arbeit anzufertigen und besonders für ihr Engagement bei der Betreuung der Arbeit sowie die Hilfsbereitschaft und die vielen wertvollen Hinweise, die zur Vollendung der Arbeit hilfreich waren.

Ein besonderer Dank gilt meiner Mutter Katrin, meinem verstorbenen Vater Edgar sowie meiner Frau Lisette, die mir immer den Rücken freigehalten haben und mich dazu animiert haben diesen Weg zu gehen.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zum **Thema „Ähnlichkeitsanalyse temporaler Patientennetzwerke auf Basis von grafenbasierten Vergleichsalgorithmen“** handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift