Otto-von-Guericke-Universität Magdeburg

Fakultät für Informatik
Institut für Simulation und Graphik

Bachelorarbeit

Bereinigung und Management klinischer Routinedaten

Autor:
Jacob Ludwig Schmidt

16. August 2020

1. Prüfender
PD Dr.-Ing. habil. Steffen Oeltze-Jafra

Fakultät für Informatik & Medizin
Otto-von-Guericke-Universität
Magdeburg
Universitätsplatz 2
39106 Magdeburg

2. Prüfende
M.Sc. Juliane Müller

Fakultät für Medizin
Otto-von-Guericke-Universität
Magdeburg
Universitätsplatz 2
39106 Magdeburg

Schmidt, Jacob Ludwig:
Bereinigung und Management klinischer Routinedaten
Bachelorarbeit, Otto-von-Guericke-Universität Magdeburg
Magdeburg, 2020.

Inhaltsverzeichnis

Abstract

In medical research, health data of human beings is essential. It forms the basis for the creation and confirmation of new hypotheses. Up to now, data to a large extent have been generated in specific studies for this purpose. A much larger amount of medical data is already being collected digitally at many healthcare institutions in the care of patients. The storage of this data takes place on proprietarily implemented databases, which are not designed for research regarding their structure, content and functionality. Accessing these data pools would drive research forward immensely.

The research group „MedDigit - Medizin und Digitalisierung“ of the Department of Neurology at the University Hospital Magdeburg is specialized in the computer-aided processing of medical data. In this thesis the data pool of the database for clinical routine data of the Clinic for Neurology of the University Hospital Magdeburg is analysed. In order to make the clinical data available to the MedDigit research group, a data integration process and a database designed for research are conceptualized and implemented. Thereby a selection, filtering, anonymisation and pseudonymisation of the data is performed. By integrating the developed processes and the new database into the system of the research group, clinical routine data of the Clinic for Neurology of the University Hospital Magdeburg is available for research for the first time.

Kurzfassung

In der medizinischen Forschung sind Daten des Gesundheitszustands von Menschen essenziell. Sie bilden die Basis für die Entwicklung und Bestätigung neuer Hypothesen. Bisher werden für diesen Zweck weitestgehend Daten in spezifischen Studien generiert. Eine weit größere Menge an medizinischen Daten wird bereits digital an vielen Institutionen des Gesundheitswesen bei der Versorgung von Patienten gesammelt. Die Speicherung dieser Daten findet auf proprietär implementierten Datenbanken statt, die in ihrer Struktur, Inhalt und Funktionsweise nicht für die Forschung ausgelegt sind. Eine Erschließung dieser Datenbestände würde die Forschung immens vorantreiben.

Die Forschungsgruppe „MedDigit - Medizin und Digitalisierung“ der Klinik für Neurologie des Universitätsklinikums Magdeburg ist auf die computergestützte Verarbeitung von medizinischen Daten spezialisiert. In dieser Arbeit wird der Datenbestand der Datenbank für klinische Routinedaten der Klinik für Neurologie der Universitätsklinik Magdeburg analysiert. Um die klinischen Daten für die Forschungsgruppe MedDigit verfügbar zu machen, wird ein Datenintegrationsprozess und eine für die Forschung ausgelegte Datenbank konzeptioniert und implementiert. Dabei findet eine Selektion, Filterung, Anonymisierung und Pseudonymisierung der Daten statt. Durch eine Integration der erstellten Prozesse und der neuen Datenbank in das System der Forschungsgruppe, stehen erstmals Daten der klinischen Routine der Klinik für Neurologie des Universitätsklinikums Magdeburg für die Forschung zur Verfügung.

1 Einleitung

1.1 Motivation

Die Digitalisierung des Gesundheitswesens führt dazu, dass immer mehr Vorgänge der Patientenversorgung, sei es die Dokumentation der diagnostischen Vorgänge, der Behandlungen oder der Untersuchungen, elektronisch aufgezeichnet werden. Krankenhäuser besitzen große IT-Infrastrukturen und Softwaresysteme - die Krankenhausinformationssysteme (KIS)-, die diese immer größer werdende Informationsmenge verarbeiten, speichern und nach Anfrage wieder zur Verfügung stellen. Die Auswertung dieser Daten hat einen enormen potenziellen Wert für die medizinische Forschung. Mit der Analyse aller im System gespeicherten Patientendaten - aus der Gegenwart wie Vergangenheit -, besteht die Möglichkeit, jede Information aus der klinischen Routine zu nutzen, um krankheitsspezifisch neue Hypothesen zu generieren oder bestehende zu verifizieren. Auch erleichtert es die Durchführung von Studien, die spezifische Aspekte untersuchen, welche nicht in den bisher vorhandenen Krankenhausdaten repräsentiert sind. Mögliche Probanden, die eventuell spezielle Kriterien erfüllen müssen, können anhand ihrer Daten aus allen Patienten in der Datenbank identifiziert werden.

Das größte Hindernis, diese Vision der medizinischen *Big Data*-Forschung umsetzbar zu machen, ist die Datenbasis selbst. In der Patientenversorgung werden viele persönliche Informationen wie Name, Geburtsdatum, Daten Angehöriger, usw. aufgenommen, die in der individuellen Behandlung nötig sind. Solche identifizierenden Daten werden im Rahmen des ethischen und rechtlichen Verständnisses des Arzt-Berufes geschützt und dürfen das Arzt-Patienten Verhältnis nicht verlassen [1]. Die Wahrung der Privatsphäre von Patienten bei einer Nutzung und Verarbeitung ihrer medizinischen Daten macht eine komplette Filterung und Anonymisierung der Daten erforderlich.

Hinzu kommt, dass die nationalen und internationalen Institutionen des Gesundheitssystems unterschiedliche Softwarelösungen [2] benutzen, sodass sich sowohl die Struk-

tur der Datenspeicherung als auch die Datenrepräsentation unterscheiden. Des Weiteren werden von medizinischen Geräten erzeugte Daten unabhängig von einheitlichen Richtlinien definiert. So folgen Hersteller und Entwickler medizinischer Geräte und IT oft ihren selbst erstellten, proprietären und nicht öffentlich zugänglichen Standards für die Datengenerierung und -speicherung [3]. Bisher gibt es nur wenige festgelegte Richtlinien, welche vorschreiben ob bzw. an welchen Standards der medizinischen Datenrepräsentation sich orientiert werden sollte [4]. Diese Heterogenität der Datenstrukturen und der Datenrepräsentationen verhindert eine effektive Vergleichbarkeit und Auswertung der vorhandenen digitalen Informationen. Sobald eines der Beiden in Bezug auf zwei Datenquellen nicht übereinstimmt, lassen sich Daten nicht mehr ohne eine vorherige Transkription für eine gemeinsame Informationsverarbeitung zusammenführen.

Das KIS „ICUData“ der Klinik für Neurologie des Universitätsklinikums Magdeburg beispielsweise speichert eingetragene Informationen in ihrer Datenbank nur in zwei Tabellen. Alle Vorgänge aus der klinischen Routine, wie Patientenpflege und Behandlung, Diagnosen, Arztbriefe, verabreichte Medikamente sowie Laboruntersuchungen werden mit krankenhausinternen Identifikationsnummern nach ihrer zeitlichen Abfolge listenförmig gespeichert. Dies hat zur Folge, dass die Extraktion von Informationen aus der Klinik-Datenbank durch einen hohen Zeit- und Kognitionsaufwand für das Suchen und Interpretieren von Dateneinträgen geprägt ist. Diese Anordnung kann systembedingt nicht verändert werden und verhindert somit eine intuitive Nutzung für Forschungszwecke.

Um die ersten Schritte der Datenanalyse der Patientendaten des Universitätsklinikums Magdeburg durchführen zu können, bedarf es deshalb einer Umstrukturierung und Neudefinition der Datenbank, in der die Datenkategorien und Datenrepräsentationen so gewählt werden, dass Informationen für Forschungsfragen einfacher extrahiert werden können.

1.2 Darstellung der Ziele und Funktionen der Arbeit

Die Forschungsgruppe „MedDigit - Medizin und Digitalisierung“ der Klinik für Neurologie des Universitätsklinikums Magdeburg untersucht Biomarkeridentifikatoren neurologischer und psychischer Erkrankungen. Durch eine digitale Analyse und automatische Prozessierung von den Bilddaten radiologischer Untersuchungen sollen Hirnstrukturen und -funktionen weiter erforscht und dokumentiert werden. Durch den Einsatz von *Radiomics, Visual Analytics* und *Deep Learning* sollen die Daten analysiert und mögliche Korrelationen zwischen Biomarkern und Krankheitsbildern aufgezeigt werden [5]. In

der Zukunft sollen auch klinische Routinedaten für identifizierende Merkmale von neurologischen Erkrankungen erfasst werden.

Das Ziel dieser Arbeit ist die Konzeption und Implementierung einer relationalen Datenbank für Patientendaten aus der klinischen Routine der Klinik für Neurologie des Universitätsklinikums Magdeburg. Dabei soll die Funktionalität der spezifischen Suche und Gruppierungen nach selbst festgelegten Kriterien hergestellt werden. Dies bedarf einer Datenintegration aus dem bestehenden Datensatz des Klinikums in das neue Datenbank-Konzept.

Es wird eine Integration der klinikinternen Vorgänge in verschiedene Relationen vorgenommen. Es bedarf einer Analyse und Aufschlüsselung der Speicherungs- und ID-Systematik des KIS „ICUData“ sowie der Repräsentation der medizinischen Daten. Dabei müssen die rein textlich verfassten Informationen teilweise in andere, bei einer Suche besser zu verarbeitende, Datentypen wie Zahlen und Wahrheitswerte umgewandelt werden. In diesem Zusammenhang wird der aktuelle Stand der Standardisierung von medizinischen Daten und Datenbanken besprochen. Der Umgang mit sensiblen medizinischen Daten wird untersucht und Möglichkeiten der Wahrung der Privatsphäre von Patienten werden konzeptioniert und implementiert.

1.3 Kapitelübersicht

In **Kapitel 2** werden grundlegende Funktionsweisen von digitaler Informationsspeicherung erklärt. Es wird eine Definition von Datenbanken (2.1) getroffen und die theoretischen Hintergründe der Struktur von Datenbanken (2.2) und des Relationenmodells (2.3) besprochen. Es wird gezeigt, wie Datenbanken durch den Einsatz von Entity-Relationship-Diagrammen (2.4) konzeptioniert werden können und wie mit Datenbanken mit der SQL-Sprache (2.5) kommuniziert werden können.

Kapitel 3 behandelt den aktuellen Stand medizinischer Datenstandards (3.1). Es werden verschiedene Datenstandards (3.1.1, 3.1.2) vorgestellt und deren Nutzung in dem Gesundheitswesen (3.2) beleuchtet.

Das **Kapitel 4** betrachtet Aspekte des Datenschutzes und der Privatsphäre von Patienten und wie diese gewahrt werden können (4.1). Anschließend wird das Mainzelliste-Programm für die Pseudonymerstellung (4.1.1) vorgestellt.

In **Kapitel 5** werden verwandte Arbeiten zu der Themenstellung dieser Arbeit vorgestellt. Das MIRACUM Projekt (5.1) und das MAGNET Projekt (5.2) werden in diesem Rahmen

diskutiert und welche Erkenntnisse sich daraus auf das Projekt dieser Arbeit übertragen lassen.

Kapitel 6 beinhaltet eine Analyse der KIS Datenbank der Klinik für Neurologie der Universitätsklinik Magdeburg (6.1) und der derzeitigen Datennutzung der Arbeitsgruppe MedDigit (6.2). Für eine Forschungsdatenbank für klinische Routinedaten werden Anforderungen definiert (6.3) und im Kontext der Bachelorarbeit analysiert (6.4).

Kapitel 7 beinhaltet die Konzeption des Datenintegrationsprozesses (7.1.1, 7.1.2) und der Forschungsdatenbank (7.1.3).

Das **Kapitel 8** behandelt die Implementierung des konzeptionierten Datenintegrationsprozesses (8.1, 8.3) und der Datenbank (8.2).

In **Kapitel 9** werden der Datenintegrationsprozesses und die Datenbank anhand der Implementierung evaluiert. Dafür werden die implementierten Funktionalitäten diskutiert und mit den gestellten Anforderungen verglichen.

Das **Kapitel 10** beinhaltet eine Zusammenfassung der Arbeit (10.1) und es wird ein Ausblick auf verschiedene Möglichkeiten der Weiterentwicklung des Projektes gegeben (10.2).

2

Grundlegende Informationen zu Datenbanken

In diesem Kapitel werden die grundlegende Funktionsweisen von digitalen Informationsverwaltungssystemen besprochen. Sie bilden die Grundlage aller modernen Datenspeicher und das Wissen über ihren Aufbau ist entscheidend für das Verständnis, die Planung und Konzeption von Datenbanken.

2.1 Definition

Eine Datenbank ist eine digitale Sammlung von strukturierten Informationen, welche die Speicherung und den Schutz vor unautorisierten Aufrufen, Manipulation und Verwaltung sowie der Prozessierung und Ausgabe von Daten erlaubt [6, 7]. Der Begriff *Datenbank* wird häufig als Synonym für *Datenbanksystem* verwendet. Ein *Datenbanksystem* besteht aus einer *Datenbank* und einem *Datenbankmanagamentsystem* (DBMS). Eine Datenbank beschreibt dabei nur die Datensätze an sich, wobei das DBMS ein Softwaremodul ist, das die Verwaltung der Datenbank übernimmt. Dazu gehört unter anderem die Integration von Daten und Überwachung der Konsistenz der Datenbank, das Ausführen von Operationen und Transaktionen, die Kontrolle sowie die Koordination der Benutzerzugriffe [8, S.7-9] [6, 7]. Wenn in dieser Arbeit von einer *Datenbank* gesprochen wird, ist immer, wenn nicht explizit anders angegeben, von einem *Datenbanksystem* auszugehen.
Der Mathematiker und Informatiker Edgar F. Codd formulierte im Jahr 1982 neun Anforderungen – auch als Regeln bezeichnet - an die Funktionalität eines DBMS [8, S.7] [9, 10]. Seine Entwicklung des relationalen Models für Datenbanken gilt als seine größte Errungenschaft und macht ihn zu einem der wichtigsten Einflussnehmer für das Anwendungsfeld der Datenbankforschung [11].

Die neun sogenannten *Codd'schen Regeln*, wortwörtlich entnommen aus „Datenbanken – Konzepte und Sprachen“ [8, S.7-8], definieren sich wie folgt:

1. ***Integration**: Die Datenintegration erfordert die einheitliche Verwaltung aller von Anwendungen benötigten Daten. Hier verbirgt sich die Möglichkeit der kontrollierten nicht-redundanten Datenhaltung des gesamten relevanten Datenbestands.*
2. ***Operationen**: Auf der Datenbank müssen Operationen möglich sein, die Datenspeicherung, Suchen und Änderungen des Datenbestands ermöglichen.*
3. ***Katalog**: Der Katalog, auch Data Dictionary genannt, ermöglicht Zugriffe auf die Datenbeschreibungen der Datenbank.*
4. ***Benutzersichten**: Für unterschiedliche Anwendungen sind unterschiedliche Sichten auf den Datenbestand notwendig, sei es in der Auswahl relevanter Daten oder in einer angepassten Strukturierung des Datenbestands. Die Abbildung dieser speziellen Sichten auf den Gesamtdatenbestand muss vom System kontrolliert werden.*
5. ***Konsistenzüberwachung**: Die Konsistenzüberwachung, auch als Integritätssicherung bekannt, übernimmt die Gewährleistung korrekter Datenbankinhalte und der korrekten Ausführung von Änderungen, so dass diese die Konsistenz nicht verletzen können.*
6. ***Zugriffskontrolle**: Aufgabe des Zugriffskontrolle ist der Ausschluss unautorisierter Zugriffe auf die gespeicherten Daten. Dies umfasst datenschutzrechtlich relevante Aspekte personenbezogener Informationen ebenso wie den Schutz firmenspezifischer Datenbestände vor Werksspionage.*
7. ***Transaktionen**: Unter einer Transaktion versteht man eine Zusammenfassung von Datenbankänderungen zu Funktionseinheiten, die als Ganzes ausgeführt werden sollen und die bei Erfolg permanent in der Datenbank gespeichert werden.*
8. ***Synchronisation**: Konkurrierende Transaktionen mehrerer Benutzer müssen synchronisiert werden, um gegenseitige Beeinflussungen, etwa Schreibkonflikte auf gemeinsam benötigten Datenbeständen, zu vermeiden.*
9. ***Datensicherung**: Aufgabe der Datensicherung ist es, die Wiederherstellung von Daten, etwa nach Systemfehlern, zu ermöglichen.*

Eine große Herausforderung in Bezug auf die Speicherung digitaler Daten stellt die Datenredundanz dar. Folgendes vereinfachtes Beispiel der Speicherung von Kundendaten,

insbesondere der Adressen, einer Produktions- und Verwaltungsfirma aus "Datenbanken - Konzepte und Sprachen"[8, S.2-3] soll diese Problematik verdeutlichen:

Die einheitliche Speicherung von Kundendaten ist für die Unternehmensverwaltung und -organisation essentiell, denn ohne eine zentrale Speicherung und Verwaltung dieser Daten werden sie redundant in verschiedenen Abteilungen vorliegen. Gibt ein Kunde eine Bestellung auf, benötigt die Versandabteilung seine Daten, um die Sendung seiner Waren durchführen zu können, und die Rechnungsabteilung, um die bestellten Waren in Rechnung stellen zu können. Sollte der Kunde zwischenzeitlich umziehen und der Firma seine neue Adresse mitteilen, muss diese Änderung an alle Abteilungen weitergeben werden, damit die Zustellung der Ware bzw. der Rechnung weiterhin gesichert ist.

Dieser Verwaltungsaufwand, der synchronisierten Aktualisierung aller Datensätze, vermag für dieses Beispiel noch überschaubar sein, jedoch vergrößert er sich immens je mehr Akteure Zugriff und Verwendung an einem Datensatz haben. Zusätzlich erzeugt eine solche Datenredundanz einen enormen Zuwachs in der Nutzung von Speicherplatz, folgend aus dem linearen Zusammenhang zwischen der Anzahl an Speicherungen eines Datums und dem Speicherverbrauch. Eine Datenbank enwickelt nach den Codd'schen Anforderungen stellt eine zentrale, gemeinsame und strukturierte Datenbasis dar, auf die alle Akteure zugreifen können. Mit einer solchen Struktur können Informationsaktualisierungen automatisch verbreitet und die eben beschriebene Problematik der Datenredundanz umgangen werden [8, S.2-4].

Dennoch kann prinzipiell nicht für alle Situationen, in denen digitale Informationen konserviert werden sollen, die Aussage getroffen werden, dass ein Datenbanksystem zum Einsatz kommen sollte. Andere Softwaresysteme, wie Textverarbeitungsprogramme oder Tabellenkalkulationsprogramme, bieten ebenfalls die Möglichkeit der Speicherung von Daten, bedienen dabei jedoch nur teilweise die Anforderungen, welche Codd [8, S.7-8] an eine Datenbank stellt. So wird das Einpflegen, Überprüfen und Manipulieren von Daten dem Benutzer überlassen, Datenredundanzen sind möglich und das Bearbeiten einer solchen Datei ist zeitgleich nur von einer einzelnen Person möglich. Grundsätzlich ist weiterhin zu beachten, wie Programme Dateien ihres Formats aufrufen und bearbeiten. So haben Microsoft Excel Dateien nach der aktuellen Version in einer 64-Bit Umgebung eine theoretisch unbegrenzte Speichergröße, können für das Öffnen auf einem Computersystem aber nicht größer als der zur Verfügung stehende Arbeitsspeicher sein. Außerdem gibt es eine Beschränkung der Zeilen auf 1.468.576 und Spalten auf 16.384 in jeder Tabelle [12]. Bei allen Programmen, deren Hauptfunktion

nicht in der langlebigen Speicherung von digitalen Daten liegt, ist die Geschwindigkeit der Öffnungs-, Bearbeitungs- und Speicherungsvorgänge dabei immer von der zugrunde liegenden Prozessor- und Speicher-Hardware des Computers abhängig, auf welchen diese stattfinden.

Es muss nach Menge und Art der zu speichernden Daten und Anforderungen des oder der Benutzer abgewogen werden, ob eine Einrichtung eines Datenbanksystems Vorteile der Kosten-, Arbeits- und/oder Zeitersparnis mit sich bringt, im Gegensatz zu dem Einsatz anderer Programme oder Lösungen.

2.2 Schematischer Aufbau und Datenunabhängigkeit

Der theoretische Aufbau einer Datenbank kann in drei verschiedene Bereiche aufgeteilt werden, die unabhängig voneinander schematisch konzeptioniert werden können. Diese Aufteilung wurde in den 1970er Jahren für die Unterstützung der Datenunabhängigkeit von der *ANSI/X3/SPARC Study Group on Database Management Systems* als „Drei-Ebenen-Schemaarchitektur"vorgeschlagen [8, S.31] [13] und ist in vielen relationalen DBMS (siehe 2.3) implementiert [14, 15]. Die Abbildung 2.1 visualisiert den Aufbau der Drei-Ebenen-Schemaarchitektur.

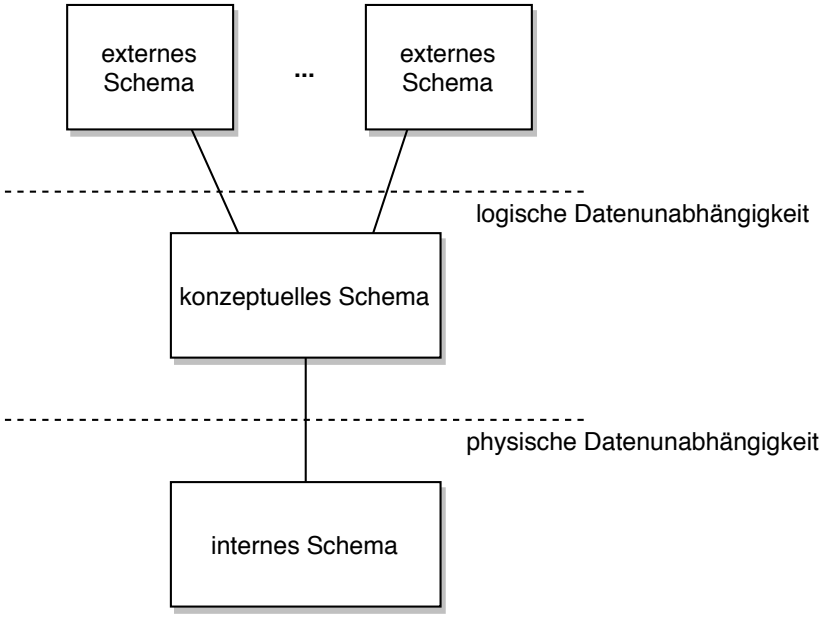

Abbildung 2.1: Drei-Ebenen-Schemaarchitektur nach Saake et al [8, S.31].

Die drei Ebenen lassen sich folgendermaßen definieren:

- Die untere Ebene – das *interne Schema* – beinhaltet die physische Struktur der Datenbank, also wie und wo die Daten tatsächlich auf einem Medium gespeichert werden. Der Aufbau der Dateien, in denen die Daten abgelegt werden, und deren Zugriffspfade werden hier festgelegt. Folglich bestehen Abhängigkeiten zwischen dem internen Schema und der zugrunde liegenden Hardwarebasis [8, S.32] [16] [17, S.269-270].
- Die mittlere Ebene – das *konzeptuelle Schema* – beinhaltet die Beschreibung der Hauptobjekte und Beziehungen der Daten. Diese Struktur wird anhand eines plattform- und implementierungsunabhängigen Datenmodells, wie zum Beispiel das Entity-Relationship-Modell (2.4) oder dem relationalen Modell(2.3), festgelegt [8, S.32] [16][17, S.269].
- Die obere Ebene – die *externen Schemata* – beinhaltet die Sichten der Nutzer oder Anwendungen auf den Datenbestand. Mit ihnen wird sichergestellt, dass Benutzer nur Informationen bekommen, für die sie berechtigt sind [8, S.32] [16] [17, S.269].

Durch die Abstraktion zwischen der physischen Datenhinterlegung, dem schematischen Aufbau der Daten und den Sichten auf bestimmte Ausschnitte der gespeicherten Daten, werden zwei beschreibbare Datenunabhängigkeiten erzeugt:

- Die *Physische Datenunabhängigkeit* beschreibt die Möglichkeit Änderungen an dem internen Schema vorzunehmen, ohne den gewählten Aufbau des konzeptuellen Schemas zu beeinflussen. Gleiches gilt auch in umgekehrter Sichtweise [8, S.31] [17, S.271].
- Die *Logische Datenunabhängigkeit* beschreibt die Möglichkeit Änderungen an dem konzeptuellen Schema vorzunehmen, ohne Funktionsweise der externen Schemata zu beeinflussen. Gleiches gilt auch in umgekehrter Sichtweise [8, S.31] [17, S.270].

Zusammengefasst soll eine Datenbank möglichst unabhängig von Veränderungen und Weiterentwicklungen ihres Umfelds bleiben, um so den effektiven Nutzungszeitraum so lang wie möglich zu halten. [8, S.30-31].

Für die Erörterung der Themenstellung dieser Arbeit und die Konzeption einer Datenbank ist vor allem das konzeptuelle Schema von Relevanz. Ein sehr verbreitetes Datenmodell ist das Relationenmodell, welches ebenfalls die Grundlage für die im Zuge dieser Arbeit erstellten Datenbank bildet.

2.3 Das Relationenmodell

Das Relationenmodell ist ein Datenmodell, welches sich auf dem schon in der Mathematik verwendeten Konzept der Relationen (lat. relatio „Beziehung“) begründet [18, S.11]. Dabei werden Daten in Beziehung zu anderen Daten gesetzt. Man erhält eine Menge von Wertepaaren, die in einer Tabelle dargestellt werden können.

Im Relationenmodell beschreiben *Relationsschemata* die zu modellierenden Objekttypen. Ein relationales Datenbankschema kann mehrere Relationsschemata beinhalten. Die Schemata beinhalten eine Menge von *Attributen*, die die Eigenschaften des zu beschreibenden Objektes darstellen [8, S.85-86]. In der Beispiel Relationen-Tabelle in Abbildung 2.2 sind dies die Namen der Spalten. Ein Relationsschema steht damit für den Kopf einer Tabelle.

Mit dem Befüllen dieser Tabelle mit Daten erhalten wir eine Instanz des Relationsschemas. Dieser Datenbereich nennt sich auch Relation. Ein Wertepaar einer Relation, d.h. eine Zeile der Tabelle, heißt *Tupel*. Da es sich bei unserer Relation um eine Menge handelt, kann jedes Wertepaar nur ein einziges Mal in einer Tabelle auftreten [8, S.85-88] [18, S.9-14]. Die Abbildung 2.2 visualisiert die Begriffsdefinitionen einer relationalen Tabelle.

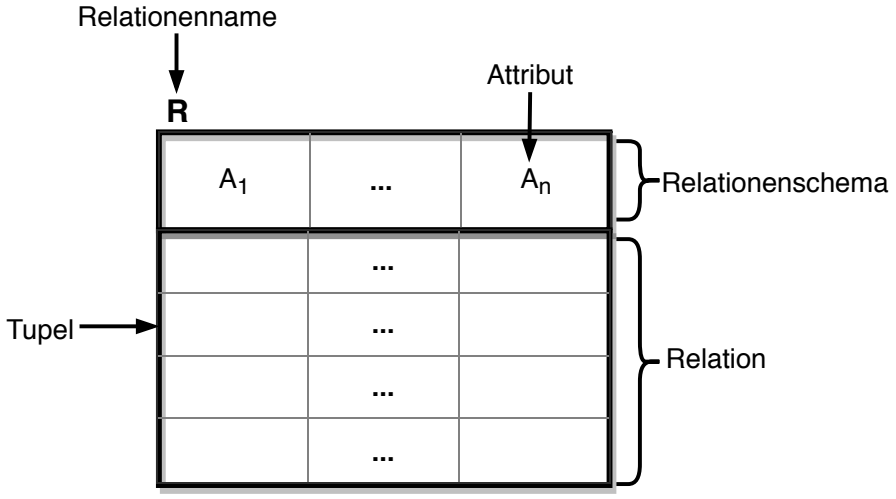

Abbildung 2.2: Veranschaulichung der relationalen Begriffe nach Saake et al [8, S.86].

Ein Datum innerhalb einer Spalte ist nun zu einem bestimmten Attribut zugehörig und muss sich in dessen festgelegten Wertebereich – auch *Domäne* genannt – befinden. Der Wertebereich wird durch den gewählten Datentyp des Attributs, meist aus der Informatik bekannte Typen, wie z.B. *Integer, Double, Char* oder *Boolean*, und durch die von dem

Entwickler vorgegebenen Einschränkungen angegeben. Wenn der Wert eines Attributs unbekannt ist, kann er durch den *Null* Wert dargestellt werden [8, S.85-88] [18, S.12-13].

Basierend auf Saake et al. [8, S.88-89, 91-94] und Studer [18, S.13-15, 17, 19-22] kann ein Relationsschema wie folgt durch mathematische Mengenregeln beschrieben werden:

Sei ein relationales Datenbankschema:= S, ein Relationsschema:= R und ein Attribut:= A. Dann sei S ein Objekt der Form

$$S := \{R_1, \ldots, R_m\},\ m \in \mathbb{N}$$

und R sei ein Objekt der Form

$$R := \{A_1, \ldots, A_n\},\ n \in \mathbb{N}.$$

Sei eine Domäne:= D und der Wertebereich eines Attributs:= $dom(A)$. Sei ein Wert:= a und eine Relation über ein Relationsschema R:= $r(R)$.
Dann sei r ein Objekt der Form

$$r(R) \subseteq \{(a_1, \ldots, a_p) | a_1 \in D_1 \wedge \ldots \wedge a_p \in D_p\}, \text{ mit } p \in N,$$

wobei

$$R = \{A_1, \ldots, A_p\},$$

$$D_1 = dom(A_1), \ldots, D_p = dom(A_p)$$

gilt.

Um Einträge (Tupel) unserer Tabelle unterscheiden zu können, brauchen wir einen *Schlüssel* für unser Relationsschema. Ein Schlüssel ist eine minimale Menge von Attributen ≥ 1, mit der man ein Objekt eindeutig identifizieren kann. Wenn die Ausprägung des/der Schlüsselattribut/e bekannt ist, kann man damit eine bestimmte Zeile der Relation ausfindig machen [8, S.90] [18, S.14-15]. Wenn mehrere Schlüssel für eine Relation vorhanden sind, muss einer von ihnen als *Primärschlüssel* ausgewählt werden. Auch wenn andere Schlüssel theoretisch Einträge identifizieren könnten, wird für die eindeutige Bestimmung einzelner Objekte in der Relation nur der Primärschlüssel verwendet. Ist nur ein Schlüssel vorhanden, ist dieser automatisch auch der Primarschlüssel [8] [18, S.15, 20-21]. Die Attribute eines Schlüssels werden auch *Primärattribute* genannt [8, 18].

Basierend auf Saake et al.[8, S.88-89, 91-94] und Studer[18, S.13-15, 17, 19-22] gilt:

Bezüglich eines Relationsschemas $R := \{A_1, \ldots, A_n\}$, $n \in \mathbb{N}$ gibt es eine identifizierende Teilmenge an Attributen $K := \{B_1, \ldots, B_q\}$, $1 \leq q \leq n$. Sei t_i dabei ein Tupel an Index i.

$$\forall\, t_i, t_j \in r(R)\left(t_i \neq t_j \Rightarrow (t_i[B_1] \neq t_j[B_1]) \vee \ldots \vee (t_i[B_q] \neq t_j[B_q])\right),\ i, j \in \mathbb{N}$$

Verallgemeinert wird diese Attributeigenschaft als eine *Funktionale Abhängigkeit* (FD, aus dem engl. „Functional Dependencies") klassifiziert. Eine funktionale Abhängigkeit einer Attributmenge X eines Relationsschemas R zu einer zweiten Menge Y aus R besagt, dass eine Ausprägung $t[X]$ eindeutig die Ausprägung $t[Y]$ identifiziert [8, S.161-163]. Genau diese Eigenschaft erfüllt auch ein Primärschlüssel: X ist in dem Fall die Menge der Primärattribute und Y die Menge aller Attribute des Relationsschemas R. Eine funktionale Abhängigkeit wird ausgedrückt durch:

$$X \rightarrow Y$$

Im Gegensatz zu einem Schlüssel darf bei einer gegeben funktionalen Abhängigkeit von X zu Y eine Ausprägung der Attributmenge X mehrmals auftreten, d.h. sie muss nicht einzigartig sein [8, S.161-163].

Mit diesem Wissen ist es möglich, einzelne Tabellen mit verschiedenen Attributen zu konzeptionieren. Objekte können durch ihre Eigenschaften beschrieben werden und durch einen zugewiesenen Primärschlüssel kann jedes eindeutig identifiziert werden.

Wie geht man aber vor, wenn ein Attribut eines Objektes ein anderes Objekt referenzieren soll. Zum Beispiel möchte man in der Tabelle *Vorgang*, mit Vorgangs-ID, Zeitpunkt, Beschreibung, auch den beteiligten Patienten speichern. Es existiert bereits eine Tabelle *Patienten*, in welcher die eindeutige Patienten-ID, der Name und das Alter gespeichert sind. Theoretisch wäre es möglich diese Patienten-ID zu kopieren und in der Tabelle *Vorgang* ein zweites Mal zu speichern, dies ist allerdings keine effektive Lösung. Eine redundante Datenspeicherung würde den Speicherverbrauch unnötig vergrößern und es wird schwierig, die Datenkonsistenz beizubehalten. Wenn in der Tabelle *Patient* die IDs verändert oder gelöscht werden, müssten diese Veränderungen manuell an allen Stellen ausgeführt werden, an denen man die IDs kopiert hat.

Mit sogenannten *Fremdschlüsseln* ist es möglich, dass ein Attribut ein anderes Objekt über dessen Primärschlüssel referenziert. In dem Wertebereich des Fremdschlüssel-Attributs liegen nur Werte, die bisher in der Relation des zu referenzierenden Objekt-Schlüssels vorhanden sind. Änderungen der referenzierten Werte (des Primärschlüs-

sels) können zu allen ihnen referenzierenden Fremdschlüsseln kaskadiert werden [8, S.93] [18, S.21-22].

Basierend auf Saake et al.[8, S.88-89, 91-94] und Studer[18, S.13-15, 17, 19-22] gilt:

Seien R1 und R2 Relationsschemata und KR2 die Attributmenge des Primärschlüssels von R2. Sei F eine Attributmenge aus R1, die R2 referenziert. Sei t ein Tupel.

$$F(R_1) \rightarrow K_{R2}(R_2), \textbf{ mit } F \subseteq R_1, K_{R2} \subseteq R_2$$

Es gilt:

$$\{t(F) \mid t \in r(R_1)\} \subseteq \{t(K_{R2}) | t \in r(R_2)\}$$

2.3.1 Weitere Integritätsbedingungen

Neben Schlüsseln gibt es noch weitere Möglichkeiten Integritätsbedingungen (engl. *Constraints*) an die Daten der Datenbank zu stellen:

- *Unique*: Einer Menge von Attributen eines Relationsschemas kann das Unique Constraint zugewiesen werden. Dies bedeutet, dass jede Kombination von Werten dieser Attribute in der Relation nur einmal vorkommen darf [18, S.19-20].
- *Not Null*: Jedem Attribut eines Relationsschemas kann das Not Null Constraint zugewiesen werden. Dies bedeutet, dass Werte dieser Attribute in der Relation nicht Null annehmen dürfen [18, S.20].

Der Primärschlüssel *Constraint* ist eine Verbindung dieser beiden Constraints. Durch die Kombination wird sichergestellt, dass mit einem Primärschlüssel Objekte eindeutig identifiziert werden können, da jeder Wert des Primärattributs nur einmal vorkommen und kein unbestimmter Wert vergeben werden darf [18, S.20-21].

2.3.2 Normalformen

Bei der Normalisierung einer relationalen Datenbank werden durch systematische Zerlegung der Tabellen durch funktionale Abhängigkeiten entstandene Datenredundanzen eliminiert und Änderungsanomalien verhindert. Änderungsanomalien können entstehen, wenn redundant vorliegende Daten verändert werden sollen: eine Update-Operation muss in diesem Fall auf alle Vorkommnisse des redundant gespeicherten Datums angewendet werden, um keine Informationsdifferenzen zu erzeugen [8, S.175].

1NF (1. Normalform):

Nur atomare Attribute

Relationsschemata dürfen nur Attribute beinhalten, die einfache Werte, d.h. keine Arrays oder Listen, abbilden. Sollte bei der Ausprägung eines Attributs z.B. eine Aufzählung gespeichert sein, muss diese aufgespalten und getrennt in der Tabelle eingetragen werden [8, S.175-176].

2NF (2. Normalform):

Elimination partieller Abhängigkeiten zwischen Schlüssel und Attributen

Ein Primärschlüssel eines Relationsschemas ist eine funktionale Abgängigkeit zwischen einer minimalen Attributmenge (auf der linken Seite), die die Ausprägung aller Attribute des Relationsschemas (die rechte Seite) bestimmt. Eine *partielle Abhängigkeit* liegt vor, wenn eine weitere funktionale Abhängigkeit einer Attributmenge zu einem Teil des Schlüssels existiert (siehe Kapitel 2.3).

Die partiellen Abhängigkeiten der nicht-Primärattribute sollen eliminiert werden. Dafür wird die Tabelle in eine weitere Tabelle zerlegt, die dem ursprünglichen zu Beginn des zweiten Normalisierungsschritt vorliegenden Relationsschema ohne die Attribute der rechten Seite der zu eliminierenden partiellen Abhängigkeit entspricht, und eine Tabelle dessen Relationsschema die Attribute beider Seiten der zu eliminierenden partiellen Abhängigkeit enthält [8, S.176-179]. Die Abbildung 2.3 veranschaulicht diesen Prozess der Elimination und Zerlegung.

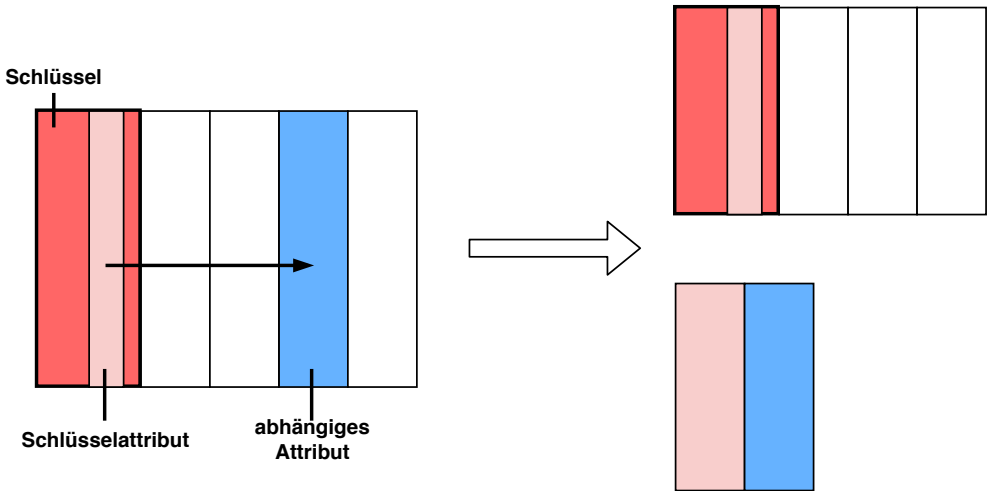

Abbildung 2.3: Überführung in die 2. Normalform durch Elimination partieller Abhängigkeiten nach Saake et al [8, S.177].

3NF (3. Normalform):

Elimination transitiver Abhängigkeiten

Eine transitive Abhängigkeit liegt vor, wenn zwei funktionale Abhängigkeiten in der Form Schlüsselattributmenge S bestimmt Attributmenge X, Attributmenge X bestimmt nicht-Primärattributmenge Y gegeben sind.

Die transitive Abhängigkeit soll eliminiert werden. Dafür wird die Tabelle zerlegt in eine Tabelle, die dem ursprünglichen zu Beginn des dritten Normalisierungsschritt vorliegenden Relationsschema ohne die Attributmenge Y der zu eliminierenden transitiven Abhängigkeit entspricht, und eine Tabelle, und eine Tabelle dessen Relationsschema die Attributmengen X und Y der zu eliminierenden transitiven Abhängigkeit enthält [8, S.179-180]. Die Abbildung 2.4 veranschaulicht diesen Prozess der Elimination und Zerlegung.

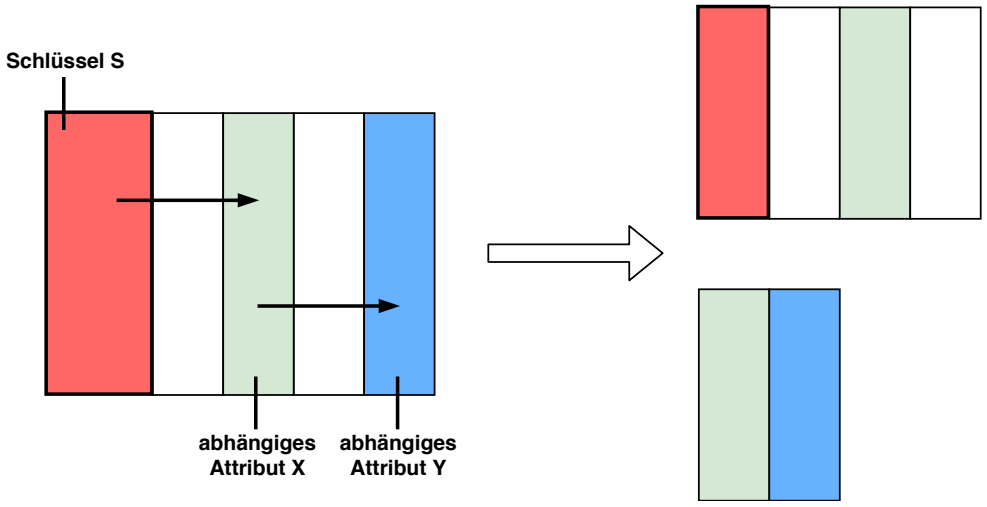

Abbildung 2.4: Überführung in die 3. Normalform durch Elimination transitiver Abhängigkeiten nach Saake et al [8, S.180].

2.4 Entity-Relationship-Modell

Das *Entity-Relationship-Modell* (ER-Modell) ist eine Notation für die grafische Modellierung von Daten und Prozessen. Es wird häufig für die Planung von Datenbanken benutzt, da es sich durch eine geringe Anzahl an Notationselementen und geringem Lernaufwand auszeichnet [19]. Es ist ein Kommunikationswerkzeug, um zwischen Anwender und Entwickler zu vermitteln [8, S.51]. Das relationale Datenmodell und das ER-Modell sind in ihrer Designsprache, -struktur und -elementen sehr ähnlich und damit

kompatibel für eine Umwandlung ineinander. Damit eine effektive Implementierung der Datenbank im Zuge dieser Arbeit gewährleistet werden kann, soll für die Planung und Visualisierung des Datenbankschemas ein ER-Modell entwickelt werden.

Das ER-Modell lässt sich in drei Grundbausteine aufteilen:

1. Entität: Ist ein definierbares Objekt, über das Daten gespeichert werden (Abbildung 2.5) [8, S.60] [20].

Abbildung 2.5: Darstellung einer Entität eines ER-Modells nach Saake et al [8, S.61].

2. Beziehung: Stellt einen Zusammenhang zwischen Entitäten her (Abbildung 2.6) [8, S.61-63] [20].

Abbildung 2.6: Darstellung einer Beziehung eines ER-Modells nach Saake et al [8, S.62].

Einer Beziehung wird immer eine Kardinalität zugewiesen. Sie gibt für die beteiligten Entitäten an, wie viele Objekte jeweils miteinander in Beziehung stehen. Eine verwendete Notation für Kardinalität ist die *[Min, Max]* Notation, die an jeder Beziehungslinie zu einer Entität ihre jeweilige Kardinalität angibt. *Min* steht dabei für die minimale Anzahl an Objekten einer Entität, die an der Beziehung teilnehmen, *Max* für maximale Anzahl an Objekten [8, S.68-75]. Es existieren:

- *1:1 Beziehung*: Ein Objekt einer Entität steht mit bis zu einem Objekt der anderen Entität in Beziehung [8, S.72].
- *1:n / n:1 Beziehung*: Ein Objekt einer Entität steht mit bis zu n Objekten der anderen Entität in Beziehung [8, S.72-73].
- *m:n Beziehung*: m Objekte einer Entität stehen mit bis zu n Objekten der anderen Entität in Beziehung [8, S.73].

Die Kardinalität ist bidirektional und kann in beide Richtungen gedeutet werden. Eine 1:n Beziehung von Entität A zu Entität B kann also gelesen werden: Ein Objekt

aus A hat eine Beziehung mit mehreren Objekten aus B oder mehrere Objekte aus B haben eine Beziehung mit einem Objekt aus A [8, S.73].

3. Attribut: Ist eine Eigenschaft einer Entität oder Beziehung. Die Einzigartigkeit der Ausprägung eines Attributes wird durch das Unterstreichen signalisiert, d.h. in allen Objekten einer Entität oder einer Beziehung kann die Ausprägung dieses Attributes nur einmal vorkommen (Abbildung 2.6) [8, S.63-65] [20].

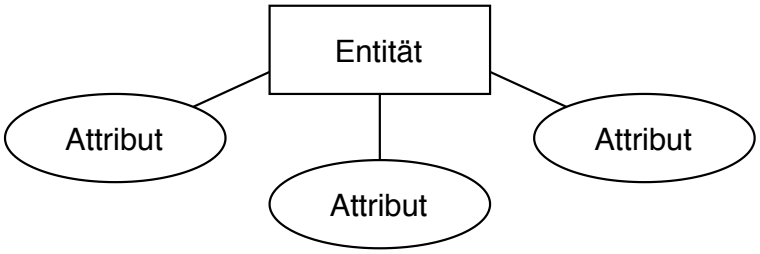

Abbildung 2.7: Darstellung eines Attributs eines ER-Modells nach Saake et al [8, S.64].

Das Zusammenspiel der Elemente lässt sich an folgendem Beispiel veranschaulichen: *Es existieren Patienten, die eine einzigartige ID, einen Namen und Alter besitzen. Es werden Behandlungen durchgeführt, deren Zeitpunkt und Vorgangsbeschreibung dokumentiert werden und an einer eindeutigen ID identifiziert werden können. Patienten und Behandlungen sollen in Verbindung gesetzt werden und zwar so, dass eine Behandlung nur einem Patienten, einem Patienten aber mehrere Behandlungen zugeordnet werden können.*

Das darauf resultierende ER-Diagramm ist in Abbildung 2.8 dargestellt.

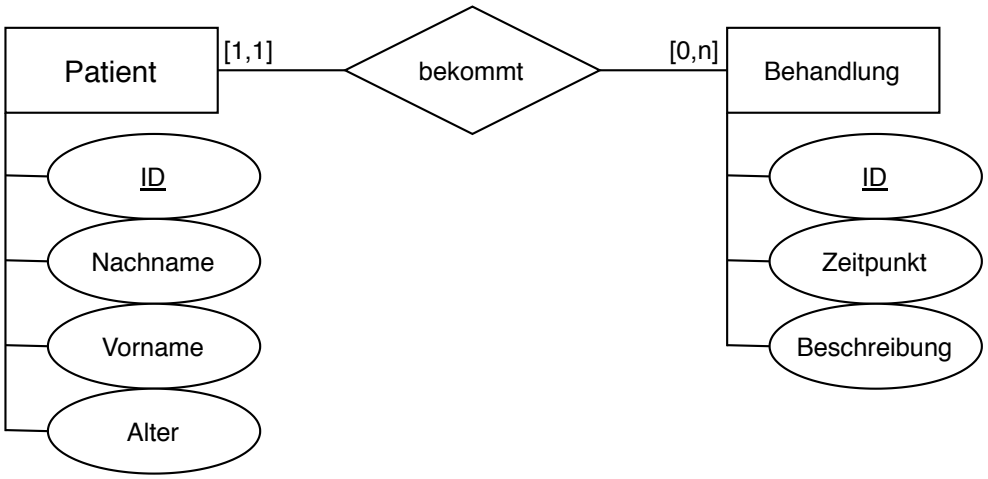

Abbildung 2.8: Beispiel eines Entity-Relationship-Diagramms.

2.4.1 Überführung eines Entity-Relationship-Modells zu einem relationalen Datenbankschema

Eine Entität lässt sich prinzipiell durch eine Tabelle, also ein Relationsschema, darstellen. Auch eine Beziehung zwischen zwei oder mehreren Entitäten lässt sich durch ein Relationsschema darstellen. Je nach Kardinalitäten einer Beziehung kann die zugehörige Tabelle auch in die Tabellen der beteiligten Entitäten integriert werden [21].

Attribute im ER-Modell lassen sich auch durch Attribute in der relationalen Datenbank darstellen. Ein Attribut einer Entität wird dabei ein Attribut des Relationsschemas der Entität. Gleiches gilt auch für Attribute von Beziehungen. Zusätzlich werden die Primärschlüssel, der an der Beziehung teilnehmenden Entitäten, Attribute des Relationsschemas der Beziehung [21].

Das Constraint eines Schlüssels wird im ER-Modell durch eine Unterstreichung der beteiligten Attribute dargestellt. Die Schlüsselmenge bildet in dem zugehörigen Relationsschema den Primärschlüssel. Für den Primärschlüssel einer Tabelle, die aus einer Beziehung entsteht, müssen die Kardinalitäten betrachtet werden. Da jeder Eintrag eines Primärschlüssels eindeutig sein muss, lassen sich an dieser Anforderung die Schlüsselmengen generieren. Bei einer 1:1 Beziehung sind die Primärschlüssel der beteiligten Entitäten Schlüsselkandidaten und einer der beiden muss als Primärschlüssel festgelegt werden. Bei einer 1:n/n:1 Beziehung wird der Primärschlüssel der Entität der n-Seite der Beziehung der Primärschlüssel. Bei einer m:n Beziehung werden beide Primärschlüssel der teilnehmenden Entitäten zu einem zusammengesetzten Primärschlüssel zusammengefügt, d.h. eine Kombination der Primärattribute muss eindeutig sein [21].

Das Ergebnis einer Überführung des ER-Diagramm aus Abbildung 2.8 zu einem relationlem Datenbankschema wird in Abbildung 2.9 veranschaulicht.

Patient

ID	**Nachname**	**Vorname**	**Alter**

bekommt

Patient_ID	**Behandlung_ID**

Behandlung

ID	**Zeitpunkt**	**Beschreibung**

Abbildung 2.9: Ergebnis einer Umwandlung eines ER-Diagramms in ein relationales Datenbankschema.

2.5 Structured Query Language

Für die Kommunikation mit der Datenbank wird eine *Structured Query Language* (SQL) benutzt. Sie bildet die Grundlage vieler relationaler Datenbankensysteme wie *PostgreSQL, Oracle Database, MySQL* oder *Microsoft SQL Server* und definiert das abstrakte (theoretische) Verhalten von strukturellen Anfragen, wie Tabellenerstellung, -bearbeitung und dem Eintragen von Daten, sowie inhaltlichen Anfragen für die Ausgabe von Daten [8, S.209] [22, S.99-100]. Festgelegt wird ihre Funktionalität in dem ISO/IEC 9075 Standard [23]. Eine SQL-Anfrage ist grundsätzlich simpel konstruiert und verständlich. Sie folgt einem natürlichsprachlichen Aufbau und ist in englischer Sprache verfasst.

Sei die Beispiel-Tabelle 2.1 gegeben.

Patienten

ID	Nachname	Vorname	Alter
23	Müller	Aaron	20
26	Lustig	Katharina	60
34	Imgarten	Manfred	30
35	Müller	Björn	50
50	Tempel	Maria	30

Tabelle 2.1: 'Patienten'.

Eine beispielhafte Datenausgabe SQL-Anfrage

```
SELECT ID, Nachname, Vorname
FROM Patienten
WHERE Nachname='Müller' AND Alter>=40;
```

erzeugt die Ausgabe-Tabelle 2.2.

ID	Nachname	Vorname
35	Müller	Björn

Tabelle 2.2: Ergebnis der Datenausgabe SQL-Anfrage: Nachname gleich „Müller“ und Alter größer gleich 40.

Eine beispielhafte Anfrage zur Eintragung eines neuen Datums

```
INSERT INTO Patienten (ID, Nachname, Vorname, Alter)
VALUES (51, 'Gottwalt', 'Thomas', 25);
```

bringt die Tabelle 2.1 in den Zustand der Tabelle 2.3.

Patienten

ID	Nachname	Vorname	Alter
23	Müller	Aaron	20
26	Lustig	Katharina	60
34	Imgarten	Manfred	30
35	Müller	Björn	50
50	Tempel	Maria	30
51	Gottwalt	Thomas	25

Tabelle 2.3: Ergebnis der SQL-Anfrage der Dateneintragung von Patient 51, „Thomas Gottwalt“ im Alter von 25 Jahren.

Wenn die Struktur, also das Relationsschema, einer Datenbank bekannt ist, ermöglicht die SQL-Sprache das Interagieren mit einer Datenbank ohne weitere Kenntnisse über zugrundeliegende Hardware, Betriebssysteme oder andere technische Details in Erfahrung bringen zu müssen. Die SQL-Kommunikation ist immer ein Bestandteil der Nutzung einer relationalen Datenbank und wird so auch für die in dieser Arbeit konzeptionierte und implementierte Datenbank essentiell sein.

3

Datenstandardisierung im Gesundheitswesen und der medizinischen bzw. klinischen Domäne

In der medizinischen Disziplin bietet eine Datenstandardisierung eine Reihe an Vorteilen. Es wird eine Vergleichbarkeit der Daten ermöglicht, die eine Interoperabilität zwischen den verschiedenen Akteuren des Gesundheitssystems herstellt. Dies bedeutet, dass Ärzte und Pflegepersonal, Forschungs- und Administrationspersonal unterschiedlicher Praxen, Stationen, Krankenhäuser oder anderen Institutionen eine klare, eindeutige Kommunikation über die Diagnosen, behandelnden Maßnahmen, Untersuchungsergebnisse und weiter wichtige Daten des Patienten haben. Administrative Aufgaben und Abwicklungen sowie die IT-Infrastruktur und Programme können vereinheitlicht werden. Der Aufwand für die Durchführung der Forschung auf ursprünglich inhomogenen/heterogenen medizinischen Daten kann stark reduziert werden [24]. Im Folgenden wird in Kapitel 3.1 auf nationale und internationale medizinische Standards der Datenrepräsentation und in Kapitel 3.2 auf die Verbreitung und Nutzung dieser Standards eingegangen.

3.1 Medizinische Standards der Datenrepräsentation

Es gibt viele nationale sowie internationale Initiativen und Aktionsbündnisse, welche Standards für die Datenrepräsentation konzeptioniert haben, sodass eine Fülle von Standards verschiedener Inklusionsreichweiten vorhanden ist. Die Standards bilden unterschiedliche Spektren der medizinischen Daten ab. Die Grenzen zwischen unterschiedlichen Bereichen sind dabei nicht immer klar definiert. So sind beispielsweise im deutschen Standard für Prozeduren („OPS", siehe Unterabschnitt 3.1.1) zusätzlich zu medizinischen Behandlungen auch Arzneimittel codiert. Ebenfalls variiert die Komplexität und Tiefe der codierten Informationen. Es gibt bisher nicht einen Gold-Standard (weder national noch international), der alle medizinischen Bereiche und Daten vereint

hat [25]. Eine Übersetzung zwischen verschiedenen Standards ist deswegen oft schwierig und aufgrund der Sensitivität der Daten ein manueller Prozess, der nur von Ärzten und medizinischem Personal ausgeführt werden kann. Es muss sichergestellt werden, dass es zu keinen Fehldeutungen oder Verlust von u.a. lebenswichtigen Informationen zwischen zwei Standard-Systemen kommt. Die Kodierung der Informationen steht in direkter Abhängigkeit zu der Implementierung der KIS und IT-Systeme der anderen Institutionen des Gesundheitswesens, die von verschiedenen Software-Anbietern bereitgestellt werden [2].

3.1.1 Medizinische Standards der Datenrepräsentation in Deutschland

Wie in Abschnitt 3.1 erwähnt, existieren bereits eine Vielzahl an verschiedenen Standards. Im Folgenden werden wichtige und etablierte Standards, die in Deutschland eingesetzt werden, vorgestellt.

ICD-10-GM (*International Statistical Classification of Diseases and Related Health Conditions*, 10. Auflage, German Modification): Ist ein Standard für Krankheiten und Diagnosen. Dieser Standard basiert auf dem ICD-10-WHO (Standard der Weltgesundheitsorganisation „WHO“), der wegen der Anforderungen und Bedürfnisse des deutschen Gesundheitswesens von dem *Deutschen Institut für Medizinische Dokumentation und Information* („DIMDI", Köln) akkommodiert wurde [26].

Aufbau: Ein ICD-10 Code besteht aus drei bis sieben Charakteren. Der erste ist ein Buchstabe, der zweite und dritte sind Zahlen. ICD-10 ist hierarchisch aufgebaut [27, 28]. Der Aufbau wird in Abbildung 3.1 veranschaulicht.

Code	**Beschreibung**
J00-J99	„Krankheiten des Atmungssystems“
J00-J06	„Akute Infektionen der oberen Atemwege“
J02	„Akute Pharyngitis“
J02.0	„Streptokokken-Pharyngitis“

Abbildung 3.1: ICD-10-GM Code Beispiel. [29]

OPS (*Operationen- und Prozedurenschlüssel*): Ist ein Standard für Behandlungen und medizinische Maßnahmen, welches von dem *DIMDI* (Köln) entwickelt wurde [30].

Aufbau: Ein OPS Code besteht aus drei bis sechs Charakteren. Der erste, zweite und dritte sind Zahlen. OPS ist hierarchisch aufgebaut [30, 31]. Der Aufbau wird in Abbildung 3.2 veranschaulicht.

Code	Beschreibung
1-10 bis 1-99	„Diagnostische Massnahmen“
1-50 bis 1-58	„Biopsie durch Inzision“
1-54	„Biopsie an Mund, Mundhöhle und Pharynx durch Inzision“
1-545	„Biopsie an anderen Strukturen des Mundes und der Mundhöhle durch Inzision“
1-545.3	„Mundboden“

Abbildung 3.2: OPS Code Beispiel. [32]

ATC/DDD (*Anatomisch-Therapeutisch-Chemisch mit definierten Tagesdosen*) Klassifikation: Ist ein Standard für medizinische Wirkstoffe und Medikamente entwickelt von der *WHO* (Genf, Schweiz) und für das deutsche Gesundheitssystem von dem *DIMDI* (Köln) angepasst [33].

Aufbau: Ein ATC Code besteht aus einem bis sieben Charakteren. Der erste, vierte und fünfte Charakter sind Buchstaben, der zweite, dritte, sechste und siebte Charakter sind Zahlen. ATC ist hierarchisch aufgebaut [33]. Der Aufbau wird in Abbildung 3.3 veranschaulicht.

Code	Beschreibung
A	„Alimentäres System und Stoffwechsel“
A02	„Mittel bei Säure bedingten Erkrankungen“
A02B	„Mittel bei peptischem Ulkus und gastrooesophagealer Refluxkrankheit“
A02BC	„Protonenpumpenhemmer“
A02BC06	„Dexrabeprazol“

Abbildung 3.3: ATC/DDD Code Beispiel. [34]

3.1.2 Medizinische Standards der Datenrepräsentation International

Im Folgenden werden wichtige und etablierte Standards, die in International verwendet werden, vorgestellt.

SNOMED-CT (*Systematized Nomenclature of Medicine – Clinical Terms*): Ist ein universeller Standard für das Gesundheitswesen entwickelt von der *SNOMED International* (London, England). Im Gegensatz zu anderen Standards zielt SNOMED auf eine Abdeckung möglichst vieler Bereiche der medizinischen Versorgung ab [35, S.11-17].

Aufbau: SNOMED besteht aus *Concepts*, *Descriptions* und *Relationships*. Ein *Concept* ist eine eindeutige Definition eines klinischen Begriffes oder Konzepts und wird durch ei-

ne eindeutige ID gekennzeichnet. Zu jedem *Concept* kann eine Menge an *Descriptions* gebunden werden. Eine *Description* ist eine textuelle Information zur weiteren Beschreibung verwendeter Synonyme des im *Concept* definierten klinischen Begriffs. Außerdem kann den *Descriptions* eines *Concepts* noch das Attribut „Preferred“ – nur einmal für jedes *Concept* - oder „Acceptable“ zugewiesen werden. Eine *Relationship* stellt einen Zusammenhang zwischen *Concepts* her. Unterschieden werden kann zwischen einer „is a“ und einer „attribute“ *Relationship*. „is a“ *Relationship* beschreibt ähnlich wie bei anderen medizinischen Datenstandards eine hierarchische Beziehung zwischen *Concepts*. „attribute“ *Relationships* sind Assoziationen und weitere definierende Charakteristiken wie Ort oder Position [35, S.17-23]. Der Aufbau wird in Abbildung 3.4 veranschaulicht.

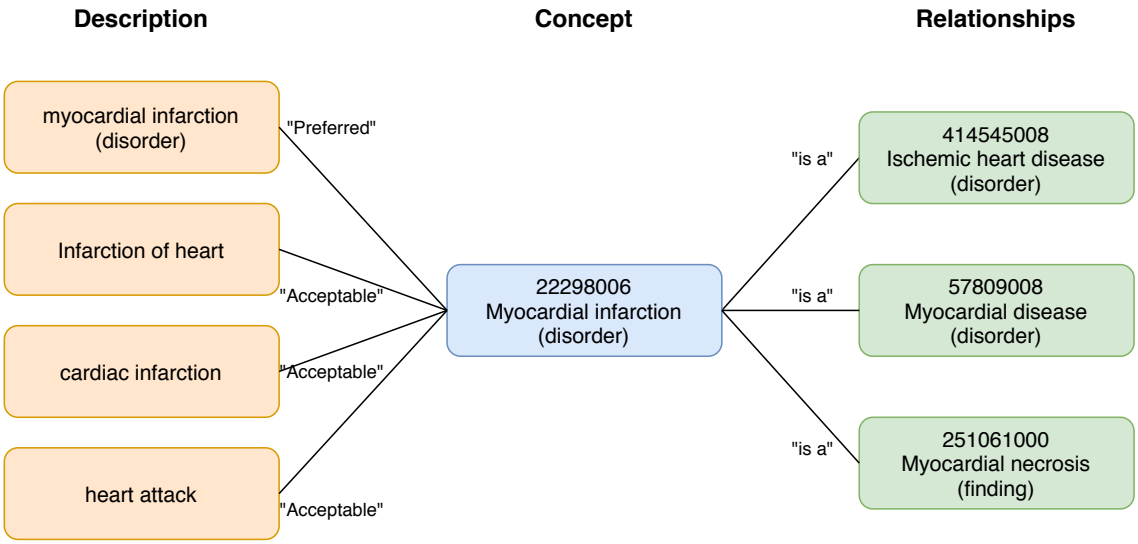

Abbildung 3.4: Beispiel einer SNOMED CT Klassifizierung nach SNOMED International [35].

CPT4 (*Current Procedural Terminology*): Ist ein Standard für medizinische Prozeduren entwickelt von der *American Medical Association* (Chicago, USA) [36].

Aufbau: CPT Codes bestehen aus fünf Charakteren. Sie können Zahlen und Buchstaben sein [36]. Der Aufbau wird in Abbildung 3.5 veranschaulicht.

Code	**Beschreibung**
10004 bis 69990	„Surgery“
40490 bis 49999	„Surgical Procedures on the Digestive System“
43020 bis 43499	„Surgical Procedures on the Esophagus“
43020 bis 43045	„Incision Procedures on the Esophagus“

Abbildung 3.5: CPT4 Code Beispiel. [37]

HCPCS (*Healthcare Common Procedure Coding System*): Ist ein Standard für medizinische Prozeduren entwickelt von *American Medical Association* (Chicago, USA) basierend auf CPT [38].

Aufbau: HCPCS wird in zwei Level oder Ebenen unterteilt. Das erste Level besteht aus den CPT Codes und beschreibt damit klinische Vorgänge und Prozesse. Das zweite Level beschreibt Produkte und Mittel, die nicht durch CPT abgedeckt sind. Ebenfalls werden Tätigkeiten, die außerhalb von Kliniken oder Praxen stattfinden, werden codiert, wie z.B. die Nutzung von Krankentransporten oder einer Notfallversorgung. Level II Codes bestehen aus fünf Charakteren, der erste ist ein Buchstabe, die restlichen sind Zahlen [39, 40]. Der Aufbau wird in Abbildung 3.6 veranschaulicht.

Code	**Beschreibung**
A0021 bis A0999	„Ambulance and Other Transport Services and Supplies“
A0021	„Ambulance service, outside state per mile, transport (Medicaid only)“

Abbildung 3.6: HCPCS Level II Code Beispiel. [41]

LOINC (*Logical Observation Identifiers Names and Codes*): Ist ein Standard für Daten aus Laboruntersuchungen und klinischen Untersuchungen entwickelt von dem *Regenstrief Institute* (Indianapolis, USA) [42, 43].

Aufbau: Ein LOINC Code besteht aus drei bis sieben Zahlen [43]. Der Aufbau wird in Abbildung 3.6 veranschaulicht.

Code	**Beschreibung**
2339-0	„Glucose [Mass/volume] in Blood"

Abbildung 3.7: LOINC Code Beispiel. [44]

RxNorm: Ist ein Standard für Arzneimittel, Medikamente und Wirkstoffe entwickelt von der *National Library of Medicine* (Bethesda, USA) [45].

Aufbau: Jedes Arzneimittel bekommt einen aus Zahlen bestehenden, eindeutigen Code - eine *RXCUI* - zugewiesen. Desweiteren gibt es sogennante *atoms*, in denen weitere Charakteristiken, Informationen und Beziehungen zu einer Arznei gespeichert werden können. Jedes atom erhält eine aus Zahlen bestehende, eindeutige *RXAUI* [45]. Der Aufbau wird in Abbildung 3.8 veranschaulicht.

Code	**Beschreibung**
18631	„Azithromycin"

Abbildung 3.8: RxNorm Code Beispiel. [46]

3.1.3 Observational Medical Outcomes Partnership (OMOP)

Dieses von der *Oberservational Health Data Sciences And Informatics* (OHDSI) entwickelte Datenmodell zielt darauf ab, viele internationale und nationale Datenstandards der an dem Gesundheitswesen beteiligten Felder in einem Datenmodell zu vereinigen. Das Ziel ist eine Harmonisierung und Vereinigung weltweiter Datenbanken, indem eine Übersetzung in ein gemeinsames Format vorgenommen wird [47].

Das *OMOP Common Data Model* (OMOP) definiert verschiedene Bereiche der Daten der Patientenversorgung: *Condition, Procedure, Measurement, Drug, Device, Observation, Visit*. Für jeden dieser Bereiche bestimmt das ODSHI ein Standard Vokabular, in das alle Konzepte translatiert werden sollen. Für *Condition* wird beispielsweise der SNOMED-CT Standard benutzt, für *Drug* der RxNorm Standard und für *Procedure* Konzepte aus den SNOMED-CT, CPT4, HCPCS und ICD10PCS Standards [48, S.63].

Das Modell lässt sich zunächst in drei Tabellen unterteilen:

- In der *concept* Tabelle befinden sich alle Konzepte beinhalteter Datenstandards. An jedes *concept* wird eine eindeutige *concept_id* vergeben und es werden Name, Ursprungs Vokabular (ursprünglicher Standard) und zugehörige Informationen gespeichert [48, S.58-65].
- In der *concept_relationship* Tabelle werden Verbindungen eines Konzeptes zu Konzepten anderer Standards festgelegt. Über ein Mapping der eindeutigen *concept_id* zu einer *concept_id* eines Synonymen Konzeptes findet hier die Übersetzung statt [48, S.65-68].
- In der *concept_ancestry* Tabelle werden Hierarchien, wie sie aus ICD-10 oder SNOMED-CT bekannt sind, hinterlegt [48, S.68-70].

Wenn nun eine Datenbasis in das OMOP Common Data Model überführt werden soll, muss zunächst geprüft werden, ob die verwendeten Datenstandards bereits in OMOP integriert sind. Sollte dies nicht der Fall sein, muss eine Integration dieses neuen Standards in das OMOP Datenmodell stattfinden, indem all seine Konzepte, Beziehungen, Hierarchien und Übersetzungen zu anderen Standards eingepflegt werden. Wenn die

verwendeten Standards sich im OMOP Datenmodell befinden, kann für jeden Dateneintrag über die Tabellen des OMOP eine Übersetzung in das jeweiligen Standard Vokabular gefunden werden und die Datenbasis wird so kompatibel mit anderen Datenbanken des OMOP Common Data Model.

3.1.4 Informatics for Integrating Biology and the Bedside (i2b2)

Das *Informatics for Integrating Biology and the Bedside* (i2b2) Datenmodell verfolgt ebenfalls das Ziel einer einheitlichen Strukturierung heterogener medizinischer Daten [49]. Ähnlich zu dem in Unterabschnitt 3.1.3 beschrieben OMOP Datenmodell soll die Nutzung dieses Modells eine Integration von Datenbeständen verschiedener Institutionen weltweiter Gesundheitssysteme in eine gemeinsame Form fördern.

Die Struktur des i2b2 Datenmodells wird durch ein Sternschema bestimmt, welches einen zentralen *Fact* durch weitere umliegende *Dimensions* erweitert [49]. Im *Fact* werden Beobachtungen oder Ereignisse eines Patienten wie Untersuchungs- und Laborergebnisse festgehalten. Eine *Dimension* ist eine Erweiterung eines solchen Ereignisses durch demografische Patientendaten, Besuchsinformationen oder Daten des behandelnden Personals. Die Abbildung 3.9 zeigt einen Ausschnitt des i2b2 Datenmodells.

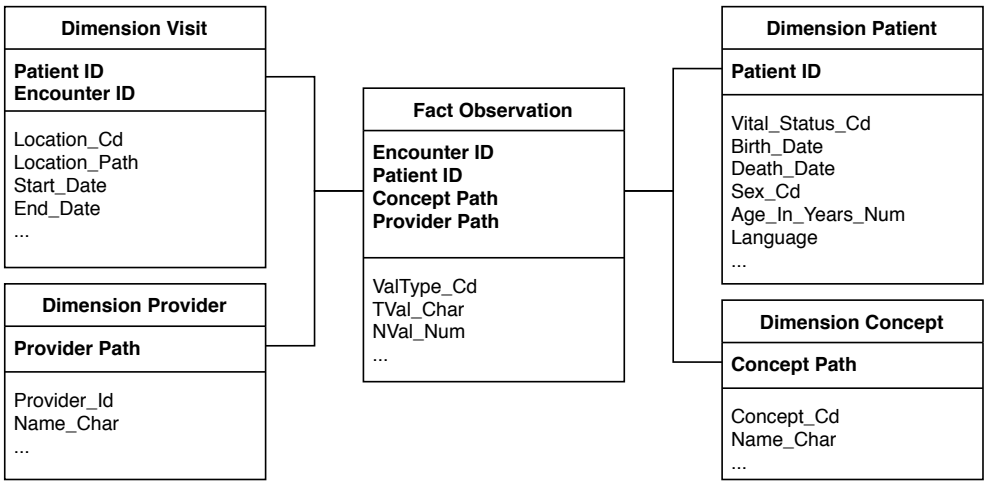

Abbildung 3.9: Ausschnitt des i2b2 Datenmodells der i2b2 Organisation [49].

Die gewählten Standards der Datenrepräsentation sind ICD-10 für Diagnosen, CPT für Prozeduren, RxNorm für Medikamente und der LOINC Standard für Laborergebnisse [49].

3.2 Nutzung von Datenstandards

Trotz einer Vielzahl an zur Verfügung stehenden Datenstandards ist ihre Nutzung in Deutschland nicht sehr verbreitet. Dies stellt ein großes Problem für eine Vereinigung verschiedener Datenquellen dar [4] [50, S.95-97, 100-101, 264]. Prof. Dr. Martin Dugas - Leiter des Instituts für Medizinische Informatik an der Universität Münster - beschreibt in seinem Artikel „Smart Medicine: Im Zeitalter von Big Data mehr als eine Vision"[51] den bisherigen Umgang mit medizinischen Daten und deren Vereinheitlichung als *Insellösungen*. Als Grund gibt er eine mangelnde Transparenz der verschiedenen Hersteller von KIS und Softwarelösungen für den Gesundheitssektor an. Eine Veröffentlichung der individuell implementierten Dokumentationsrichtlinien und Formularen, sowie deren Datenspeicherung, finde dabei nicht statt und verhindere so eine gemeinschaftliche Erarbeitung von einheitlichen Datenmodellen.

Seine Aussagen werden weiterhin von anderen Wissenschaftlern aus dem Feld der Medizininformatik, wie z.B. Prof. Dr. Hans-Ulrich Prokosch - Lehrstuhlinhaber für Medizinische Informatik an der Friedrich-Alexander Universität Erlangen Nürnberg und Leitung des „MIRACUM“ Konsortiums -, gestärkt. Gemeinsam machten sie bereits 2014 mit ihrer Arbeit „Freier Zugang zu Dokumentationsformularen und Merkmalskatalogen im Gesundheitswesen. Memorandum Open Metadata“[3] darauf aufmerksam, dass ein Großteil der für Studien und in Informationssystemen des Gesundheitssystems genutzten Informationsformulare nicht öffentlich verfügbar seien. Eine vertragliche Unzulässigkeit der Veröffentlichung von Inhalten der Datengewinnung und -verarbeitung solcher Softwaresysteme lässt sich unter wirtschaftlichen Gesichtspunkten und dem Interesse der Entwickler der Plattformen, ihr geistiges Eigentum zu schützen und zu bewahren, nachvollziehen. Jedoch verhindert dies eine organische Entstehung oder Nutzung von Standards für die Erhebung und den Umgang mit medizinischen Daten.

Auf internationaler Ebene gibt es ebenfalls Bestrebungen für eine verantwortungsvolle Teilung von Daten der Akteure im Gesundheitswesen. So wurde im „The Lancet“ Journal [52] und im „The New England Journal of Medicine"[53] von dem „International Committee of Medical Journal Editors“ (ICMJE) ein Kommentar veröffentlicht, welcher dazu aufruft, zugrundeliegende Daten medizinischer Studien anonymisiert verfügbar zu machen.

Auch der Deutsche Ethikrat stellt in seiner Stellungnahme zu „Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung“[50, S.95-97, 100-101, 264] eine Heterogenität der Datenerfassung in Deutschland fest und empfiehlt Verein-

heitlichung und Standardisierung von medizinischen Datenmodellen unter Berücksichtigung der Datenschutzrechte und der Privatsphäre von Bürgerinnen und Bürgern.

Eine Ausnahme bilden Standardisierungen, deren Nutzung gesetzlich vorgeschrieben ist oder welche für die Administration, Verrechnung und Abrechnung entstehender Kosten im Gesundheitswesen eingesetzt werden [4]. Nach § 17b des Krankenhausfinanzierungsgesetztes wird in Deutschland das Vergütungssystem *German Diagnosis related groups* (G-DRG) genutzt, um stationäre oder ambulante Behandlungen eines Krankenhauses abzurechnen [54, 55]. ICD-10-GM und OPS bilden dabei die grundlegende Kodierung der abrechenbaren Leistungen für das G-DRG-System und sind deswegen in den Arbeitsablauf von Krankenhäusern integriert [54, 55].

Das deutsche Bundesministerium für Bildung und Forschung setzt sich verstärkt für eine Vernetzung und Digitalisierung von Akteuren des Gesundheitswesens ein [56]. Bis 2021 fördert es unter der „Medizininformatik-Initiative“ (MI-I) Forschung und Entwicklung einer deutschlandweiten, gemeinschaftlichen digitalen Infrastruktur für die Harmonisierung und Integration biomedizinischer Daten mit ungefähr 160 Millionen Euro [57]. Zu der jetzigen Zeit sind vier Konsortien – „DIFUTURE“, „HiGHmed“, „MIRACUM“ und „SMITH“ - beauftragt, Konzepte und Produkte zu entwickeln, die einen Daten- und Wissensaustausch ermöglichen und befördern [58, 59].

3.3 Fazit für eine Datenbank der Arbeitsgruppe MedDigit

Die in diesem Kapitel besprochenen Aspekte der Datenstandardisierung in der medizinischen Domäne zeigen eine Reihe an Vorteilen für die Nutzung von medizinischen Daten in der Analyse für Forschungszwecke. Die Komplexität der Semantik verschiedener medizinischer Therme bedeutet allerdings auch, dass eine Übersetzung unterschiedlicher Begriffe oder Sprachsysteme eine schwierige und langwierige Tätigkeit ist, bei der medizinisches Fachpersonal unabdingbar ist. Der resultierende Aufwand übersteigt die Möglichkeiten, die in dieser Bachelorarbeit zur Verfügung stehen, und daraus folgt, dass in der in dieser Arbeit zu konzeptionierenden und implementierenden Datenbank keine neuen Datenstandards eingeführt oder in bestehende translatiert werden.

4

Anonymität und Privatsphäre

Bei der Verarbeitung medizinischer Daten müssen ethische und rechtliche Aspekte des Datenschutzes beachtet werden [1]. Die Privatsphäre und Verschwiegenheit spielen eine wichtige Rolle in dem Verhältnis zwischen medizinischem Personal und Menschen, die bei ihnen Behandlung ersuchen und bekommen. Für eine effektive Behandlung ist das Vertrauen des Patienten gegenüber dem Arzt essenziell. Erst dieses Vertrauen bildet die Basis für den uneingeschränkten Austausch von Informationen. Der Patient geht in der Annahme, dass persönliche Informationen in seinem eigenen Interesse und nicht gegen ihn verwendet werden. Die Diskretion ist Teil des ethischen Selbstverständnisses medizinischen Personals im Sinne der ärztlichen Schweigepflicht [1].

Für Forschungsarbeiten in der medizinischen Domäne sind Daten des menschlichen Gesundheitszustands grundlegend für die Erstellung, Überprüfung und Bestätigung neuer Hypothesen von Bedeutung. Eine Veröffentlichung medizinischer Aufzeichnungen würde diesen Prozess immens vorantreiben. Es entsteht jedoch ein ethisches Dilemma zwischen Privatsphäre und Rechten von Individuen zu dem Interesse der Gesamtheit der Gesellschaft an wissenschaftlichen und medizinischen Fortschritten. Eine Überwindung dieser Problematik ermöglicht es, ein weites Spektrum an wissenschaftlichen Ansätzen über eine neue Datenquantität zu verfolgen.

4.1 Anonymisierung und Pseudonymisierung

Eine Lösung dieses Konfliktes liegt darin, einen Weg zu finden, der eine Nutzung medizinischer Daten bei gleichzeitiger Erhaltung des Bedürfnisses auf die Privatsphäre ermöglicht. Die Problematik ergibt sich, wenn die medizinischen Daten von ihren erzeugenden Individuen losgelöst werden, sodass persönliche Informationen nicht das Arzt-Patienten-Verhältnis verlassen. Durch eine Anonymisierung aufgezeichneter medizinischer Daten könnte die Privatsphäre von individuellen Patienten gewahrt werden. Eine

Elimination persönlicher Informationen verhindert eine Identifizierung einzelner Personen. Dadurch bleibt das Vertrauensverhältnis zwischen Behandelnden und Patienten erhalten und private medizinische Angelegenheiten bleiben der Öffentlichkeit verborgen.

Eine Erstellung von Pseudonymen als Ersatz der Personen macht es möglich, Daten an ein abstraktes Individuum zu binden und eine Nachverfolgbarkeit wiederherzustellen, sodass ein Anonymisierungsvorgang von medizinischen Daten keine substantiellen Nachteile und Informationsverluste nach sich zieht.

4.1.1 Mainzelliste

Die Mainzelliste ist ein von der Universität Mainz entwickeltes Programm für die Erstellung und Verwaltung von Pseudonymen im Rahmen der Verarbeitung medizinischer Daten. Das Programm vereint eine Kommunikation von Patientendaten über ein REST-Interface - einem Standard für eine simple Client-Server Kommunikation -, eine Erzeugung von Pseudonymen und die Speicherverwaltung in einer angeschlossenen Datenbank [60, 61].

Die Mainzelliste wird auf einem Webserver gestartet und mit HTTP-Anfragen kann der Nutzer Anfragen stellen, Patienten hinzuzufügen, zu editieren oder auszulesen. Bei der Initialisierung können frei wählbar Attribute der Mainzelliste definiert und deren Eingaben konfiguriert werden. So können beispielsweise Namen und Adressen als textuelle Eingaben definiert werden und Geburtsdaten als numerische Eingaben, die nur in einer gültigen Kombination akzeptiert werden sollen. Bei der Eintragung eines Patienten - einer *addPatient*-Anfrage - werden Ausprägungen dieser Patientendaten an die Mainzelliste übergeben. Das Programm führt nun einen Matching-Algorithmus mit bereits gespeicherten Patienten aus - gibt es genug übereinstimmende Werte entsteht ein Match und es wird das schon bestehende Pseudonym ausgeben, in der Annahme beide Patienten sind ein Individuum. Diese Eigenschaft ist dafür geeignet Personen über Datensätze unterschiedlicher Struktur und Qualität zu verfolgen [61]. Wird kein Match gefunden, wird ein neues Pseudonym generiert.

Die Pseudonyme entstehen aus einem Algorithmus, der aus einem internen Zähler und einem selbst wählbaren Schlüssel ein Pseudonym errechnet[62]. Bei gleich bleibendem Schlüssel heißt dies auch, dass bei einem Zurücksetzen der Mainzelliste, die gleichen Pseudonyme in der gleichen Reihenfolge wieder erzeugt werden würden, unabhängig von der Eingabe. Dies bedeutet, dass eine nach einer bestimmten Konfiguration erstellte Instanz einer Mainzelliste für eine kohärente Pseudonymerstellung über den kom-

pletten Datensatz hinweg verwendet werden muss. Ein Pseudonym kann aus bis zu 30 bits bestehen, sodass die maximale Anzahl an zur Verfügung stehenden Pseudonymen 2^{30} - ca. 1.000.000.000 - ist [62]. Eine HTTP- Auslese-Anfrage - eine *getPatient*-Anfrage - ermöglicht es, mit der Eingabe eines Pseudonyms die Patientendaten, falls vorhanden, wieder auszulesen.

Die eingegebenen Daten der Attribute und das erzeugte Pseudonym werden in einer an die Mainzelliste angeschlossenen relationalen Datenbank gespeichert.

Der Einsatz eines Pseudonymisierungsprogramms wie Mainzelliste ermöglicht es den in diesem Kapitel angesprochenen Konflikt zwischen der Privatsphäre des Patienten und dem Nutzen persönlicher medizinischer Daten für Forschungszwecke zu überwinden, indem es neue abstrakte Individuen schafft, die den Platz von real identifizierbaren Personen einnehmen. So ist eine vollumfängliche Nutzung der Daten in der Wissenschaft bei gleichzeitiger Wahrung der Identitäten aller Beteiligten möglich.

5 Verwandte Arbeiten

Die Erschließung klinischer Datenbestände für den Einsatz in der medizinischen Forschung stellt den zu untersuchenden Tatbestand in dieser Arbeit dar. Das MIRACUM Projekt (Abschnitt 5.1) ist ein vollumfänglicher Versuch der Erstellung einer national vernetzen Datenbank klinischer Daten, das einen großen Aufschluss über die verschiedene Aspekte und Probleme der Planung und Implementierung dieser gibt. Anschließend wird eine weitere Perspektive auf das Thema der Arbeit mit den Erfahrungen des MAGNET-Projekts (Kapitel 5.2) geliefert.

5.1 Medical Informatics in Research and Care in University Medicine (MIRACUM)

Das *Medical Informatics in Research and Care in University Medicine* (MIRACUM) Konsortium verfolgt seit 2018 unter Förderung des Bundesministeriums für Bildung und Forschung das Ziel eine nationale Vernetzung biomedizinischer Datenbanken der Institutionen des deutschen Gesundheitswesens zu schaffen [63]. Das Universitätsklinikum Magdeburg ist ein aktives Mitglied dieses Konsortiums und trägt zu dessen Entwicklung bei. Damit ist dieses von besonderem Interesse in Bezug auf die Themenstellung dieser Arbeit. Für die Erstellung einer zentralen Datenbank benutzen sie eine Reihe an Open Source Software Werkzeugen, dessen Zusammenstellung unter dem *MIRACOLIX* Stichwort gehandelt wird [64]. Die Kategorien des MIRACOLIX Pakets bilden dabei den theoretischen Prozess der Integration einzelner, lokaler Datenquellen in ein vernetztes, nationales Datenverzeichnis ab und lauten wie folgt [64]:

- ***Data Repositories***:
  Für die Speicherung von Daten werden verschiedene Lösungen vorgeschlagen. Unter anderem die Programme und Datenmodelle der *i2b2 tranSMART Foundation* (siehe Unterabschnitt 3.1.4), der *OHDSHI* (siehe Unterabschnitt 3.1.3) und des

XNAT Projekts. Diese Modelle stellen unterschiedliche Möglichkeiten der Strukturierung biomedizinischer Daten dar, die getrennt voneinander konfiguriert und implementiert werden.

- ***Project & Trial Management***:
  Es soll ein Verzeichnis für alle Studien und Projekte einer Institution konzeptioniert werden, die alle Informationen und Metadaten an einem Ort vereinen soll. Außerdem soll eine Möglichkeit implementiert werden, Studienvorschläge einzureichen und diese zu überprüfen.

- ***Data Protection***:
  Es wird ein Werkzeug für ein vernetztes *ID-Managament* entwickelt, das zu einem späteren Zeitpunkt mit der Mainzelliste-Software zusammengeführt werden soll. Das ID-Management wird dafür genutzt, Individuen über verschiedene Datenbanken hinweg, wenn z.B. ein Patient Behandlungen in mehreren Krankenhäusern erfahren hat, zu identifizieren. Für die Anonymisierung von Patientendaten soll je nach Anwendungsfall ein Programm zur Verfügung gestellt werden. Es soll ein Werkzeug geschaffen werden für die Erstellung und Verwaltung der Zustimmungen von Patienten zu der Verwendung ihrer Daten für bestimmte Forschungszwecke.

- ***IT-Infrastructure***:
  Es soll eine Software-*Pipeline* erstellt werden, die eine Entwicklung und Weiterentwicklung der MIRACOLIX Werkzeuge ermöglicht.

- ***Data Integration Tools***:
  Es sollen *ETL*-Werkzeuge (*Extract, Transform, Load*) für die Extraktion, Übersetzung und Integration von Daten in gemeinsame Verzeichnisse bereitgestellt werden. Für eine Analyse und Integration von Freitext-Daten wie z.B. Arztbriefen sollen *Natural-Language-Processing*-Werkzeuge bereitgestellt werden. Die gewählten Datenrepräsentationen der medizinischen Informationen werden in einer *Meta Data Repository* (MDR) festgehalten.

Die hierbei entstehende Datenbank soll national standardisiert und vernetzt, jedoch nicht physisch zentralisiert gespeichert sein [63]. D.h., die sensiblen medizinischen Daten bleiben zunächst lokal an ihren Entstehungsorten gespeichert und werden nur für gezielte Anfragen anonymisiert zusammengeführt. Anfragen bestimmter Daten werden an eine zentrale, verwaltende Einheit des Projekts geschickt, die die Validität der Anfrage überprüft und danach entsprechende Datenanfragen an die einzelnen lokalen Da-

tenbanken der Krankenhäuser und teilnehmenden Institutionen verschickt [65]. Für die Erstellung und Verwaltung der lokalen Datenbank richtete jedes Mitglied des Konsortiums ein *Datenintegrationszentrum* (DIZ) ein. Die DIZ sind für den Einsatz, die Entwicklung und Weiterentwicklung der MIRACOLIX Softwaretools zuständig [63]. Das MIRACUM Projekt bietet so potenziell eine allumfassende Lösung für ein vernetztes und digitales Gesundheitswesen, das große Vorteile für die medizinische Forschung verspricht. Die vorgeschlagenen Modelle, Programme und IT-Strukturen stellen effektive und leistungsfähige theoretische Lösungsansätze für den Aufbau einer Datenbank klinischer Routinedaten dar. Jedoch werden nur wenige Aussagen über den ersten Prozess der Datenintegration und -harmonisierung getroffen, der vorgeben würde, wie das geplante gemeinsame Datenverzeichnis erreicht werden kann.

Um eine Vergleichbarkeit zwischen den biomedizinischen Daten der Konsortiums Partner herzustellen, müssen die Datenbestände in eine gemeinsame, standardisierte Struktur und Repräsentation (siehe Abschnitt 3.1) gebracht werden. Die Grundvoraussetzungen sind für diesen Prozess sehr verschieden, da mehrere KIS mit unterschiedlichen Implementierungen und Infrastrukturen im Einsatz sind. Hinzu kommt, dass durch unterschiedlichen Untersuchungsverfahren und -geräte und unterschiedlichen Dokumentationsverfahren und -richtlinien jede einzelne Datenbasis einzigartig in Struktur, Aufbau und Inhalt ist. Die Aufgabe der DIZ besteht nun unter anderem darin, die lokalen Datenbanken zu sichten und anhand eines gemeinsames Datenmodells zu restrukturieren [63].

Die Translation der individuellen Datenbestände stellt einen ähnlichen Vorgang dar, welcher in Abschnitt 3.1 bei der Überführung medizinischer Datenstandards beschrieben wird. Es bedarf an medizinischem und informationstechnischem Personal, das mit dem Quellsystem – der ursprünglichen Datenbank des KIS – sowie dem Zielsystem – einer Datenbank nach MIRACUM Vorgaben – vertraut ist, um sowohl eine Umformung der Datenstruktur als auch eine Umformung der Datenrepräsentation durchführen zu können.

Im Rahmen des im April 2020 getagten Symposiums der Mitgliedern des MIRACUM Konsortiums wurden Einblicke in den aktuellen Fortschritt der projektbezogenen Arbeiten gewährt. Eine prototypische Datenintegration im Zuge einer nationalen Vernetzung - über alle Partnerinstitutionen des Konsortiums hinweg - wurde anhand eines gemeinsamen i2b2-Verzeichnisses durchgeführt [66]. Dabei wird an jedem Standort lokal eine i2b2 Datenbank eingerichtet. Über einen zentralen *Broker* werden Anfragen an das zentrale Verzeichnis an die einzelnen lokalen Datenbanken vermittelt.

Die vorhandenen integrierten und harmonisierten Daten beschränken sich bisher auf Datensätze basierend auf dem § 21 des Krankenhausentgeltgesetz (KHEntgG) [65]. Diesem Gesetzt entsprechend müssen Patienten-versorgende Institutionen jährlich Leistungsdaten an eine Datenstelle des *Institut für das Entgeltsystem im Krankenhaus* (InEK) übersenden, welche unter anderem demografische Informationen und Daten über Diagnosen – in Form von ICD-10-GM Codes (3.1.1) – und Behandlungsvorgänge – in Form von OPS Codes (3.1.1) – beinhalten [67]. Dieser Datensatz nach § 21 liegt in einem bereits vorgegebenen Format vor, welchem alle einsendenen Institutionen folgen. Die hier hinterlegten Daten sind direkt nutzbar, da eine Vergleichbarkeit und Vereinbarkeit der Informationen gegeben ist, ohne dass eine Form der Restrukturierung oder Translation vorgenommen werden muss. Es können so erste medizinische Daten in ein gemeinsames Datenverzeichnis aufgenommen werden, ohne dabei auf die eigentlichen Datenbanken der Krankenhäuser zurückzugreifen und damit verbundene Integrations- und Harmonisierungsarbeiten durchführen zu müssen.

Des Weiteren wurde ein webbasiertes Forschungsportal eingerichtet, dass später den Zugang zu dem zentralen Datenverzeichnis ermöglichen soll. Dabei müssen Anfragende einen Antrag stellen, in dem sie im Detail darstellen, welche Daten sie erhalten wollen und wofür diese Daten verwendet werden sollen. Die Anträge werden von einer bevollmächtigten Leitung bearbeitet und auf Korrektheit überprüft [65].

Das MIRACUM Projekt gibt wichtige Punkte vor, die bei der Planung und Konzeptionierung einer Datenbank für medizinische Daten beachtet werden müssen. Eine Analyse aller Aspekte ist ein wichtiger Bestandteil der Projektplanung und besonders bei einem Vorhaben dieser Tragweite unabdingbar. Einzelne Vorgänge der Datenprozessierung müssen in Bezug zueinander konzeptioniert werden, um einen effektiven und zielgerichteten Prozess zu erhalten. Grundsätzlich ist bei diesem Projekt aber anzumerken, dass das Ziel einer gemeinsamen Big Data Nutzung von medizinischen Daten nur erreicht werden kann, wenn Quelldatenbanken, wie die des Universitätsklinikums Magdeburg, in einem replizierbaren und automatisierbaren Vorgehen erschlossen werden können. Eine manuelle Verarbeitung und Integration sollte in Anbetracht der immensen Datenbestände der Krankenhäuser aus der Vergangenheit, Gegenwart und viel wichtiger noch der Zukunft - mit einer weiter voranschreitenden Digitalisierung des Gesundheitswesens und der Patientenversorgung - in jedem Fall umgangen werden.

Die seit 2017 veröffentlichten Arbeiten des MIRACUM-Projektes zeigen Konzepte und Implementierungen in Bereichen des Datenmanagements, der Qualitätssicherung, des Einwilligungsmanagaments, der Forschungsanbidung und der Vernetzung, welche al-

lesamt Lösungen bieten, die auf eine grundlegende Datenintegration aufbauen [63, 64, 65, 66]. Eine grundlegende Datenintegration befindet sich ausgehend ihrer Publikationen [68] noch in der Konzeptionsphase und ist noch nicht implementiert.

Die Präsentation der in 3 Jahren erreichten Fortschritte des Projekts zeigen einen Fokus, der nicht auf dem wichtigsten Element des Projekts liegt: die elementare Datenintegration der Quelldatenbanken. Denn sie bildet das Fundament auf dem alle anderen MIRACUM Funktionalitäten logisch aufbauen.

5.2 Die Melbourne East Monash General Practice Database (MAGNET)

Die *Melbourne East Monash General Practice Database* (MAGNET) ist ein im Jahr 2013 gestartetes Projekt für die Erstellung einer Datenbank für Informationen der Patientenversorgung allgemeinmedizinischer Praxen. Insgesamt tragen 50 Institutionen der „Inner East Melbourne Medicare Local“ (IEMML) Organisation aus der gleichnamigen Region Australiens elektronische Gesundheitsdaten von über einer Millionen Patienten zu einem gemeinsamen Forschungsaufwand zusammen. Die Sammlung soll erstmals eine Möglichkeit der Analyse einer bisher unzugänglichen Menge an medizinischen longitudinal Daten aus der Patientenversorgung bereitstellen und so neue Erkenntnisse in der Medizin und der gesundheitlichen Versorgung gewonnen werden.

Die Daten unterliegen, ähnlich zu der Situation des MIRACUM-Projekts in Deutschland, verschiedenen Strukturierungen und Repräsentation aufgrund der Nutzung unterschiedlicher Software-Lösungen zur Eintragung und Speicherung. Die Extraktion und Harmonisierung der Informationen wird mithilfe des *GRHANITE* Softwaretools durchgeführt, das von dem „Rural Health Academic Centre“ (RHAC) der Universität Melbourne entwickelt und verbreitet wird. Das GRHANITE Programm extrahiert die zu sammelnden medizinischen Daten der Praxen, anonymisiert sie, verschlüsselt sie und, im Gegensatz zu dem im MIRACUM-Projekt gewählten dezentralen Speicherung, verschickt diese Daten an eine zentrale MAGNET-Datenbank. Die Datenbank ist ähnlich zu dem *i2b2*-Datenmodell (siehe Unterabschnitt 3.1.4) in einer *star structure* angeordnet, die besagt, dass ein zentraler Fakt von bildlich gesprochen außen liegenden Informationen erweitert wird. Eine solche Anordnung kann an dem Beispiel eines Arztbesuches veranschaulicht werden: Der „Besuch“ selbst ist der Fakt. Die genaue „Zeit“, der „Ort“, der behandelnde „Arzt“ usw. stellen zugehörige Informationen dar, die separat gespeichert

werden und bildlich in der Struktur sternenförmig um den Fakt herum angeordnet werden können.

GRHANITE stellt eine Datenkompatibilität aus den in Krankenhäusern und Praxen genutzten IT-Systemen her und löst die grundlegende Kernproblematik einer dezentralisierten Digitalisierung (siehe Abschnitt 3.2), mit der sich auch das MIRACUM-Konsortiums auseinandersetzt. Die Entwicklung dieser *Middleware* als Übersetzer zwischen vereinzelten Institutionen des Gesundheitswesen und einer zentralen Datenbank findet allerdings schon seit 2006 statt und benötigte 5 Jahre, um eine erste Reihe an in Australien verwendeten digitalen Dokumentationssystemen zu unterstützen. Trotz allem gibt es nach weiteren Jahren der Entwicklung immer noch Probleme der Extraktion nicht-standardisierter Informationen und die Verwendung verschiedener Kodierungen und Terminologien kann eine Vergleichbarkeit von Daten verhindern. Die Länge der Zeit, die benötigt wurde, bis zu der Implementierung einer funktionsfähigen aber verbesserungswürdigen Lösung, gibt Aufschluss über die Komplexität dieser Datenintegrations- und -harmonisierungs-Problematik und gewährt eine weitere Perspektive auf die Tragweite des weiteren Vorgehens und der Entwicklungsarbeiten des MIRACUM-Projekts.

5.3 Schlussfolgerung

Aus dem in Abschnitt 5.1 beschriebenen MIRACUM Projekt könnten theoretisch wegweisende Erkenntnisse der Konzeption einer vernetzen Datenbank sensibler medizinischer Informationen entnommen werden. Ohne eine Erschließung des ursprünglichen Datenbestandes aus der Datenbank des KIS der Universitätsklinik Magdeburg sind diese Vorgehensweisen nur von bedingtem Nutzen für die Ausarbeitung einer neuen Datenbank in dieser vorliegenden Bachelorarbeit. Basierend auf den bisherigen Erkenntnissen dieses Kapitels ist es erforderlich eigene Ansätze zu formulieren, um eine Datenintegration und -harmonisierung zu erreichen.

6 Anforderungsanalyse

In diesem Kapitel werden die Anforderungen der Arbeitsgruppe MedDigit an eine Datenbank für klinische Routinedaten dokumentiert. Dafür findet zunächst eine Analyse der zur Verfügung stehenden Daten aus dem KIS der Klinik für Neurologie des Universitätsklinikums Magdeburg statt.

6.1 Aktueller Zustand der klinischen IT-Struktur

Die Klinik für Neurologie des Universitätsklinikum Magdeburg verwendet für die digitale Dokumentation der Patientenversorgung das KIS „ICUData“ als primäres IT-Framework und zentrales Verwaltungssystem. Dieses System bietet einen hohen Anpassungsgrad in Bezug auf das Anlegen von Formularen der Dateneintragung. So kann dazu berechtigtes Personal die Art der Informationen, ihren Datentyp und die Art der Einpflege in ausgewählte Formulare oder Dialogfenster festlegen und damit aktiv den Prozess der Datengeneration beeinflussen. Jede Aktion oder Vorgang, welcher in das System übertragend wird, wird in dem KIS in einer *Oracle Database* gespeichert. Eine Oracle Database ist ein relationales Datenbanksystem mit welchem über SQL kommuniziert werden kann.

Für einen benutzerfreundlicheren Zugriff auf die relationale Datenbank existiert ein Softwaretool, das in einer grafischen Benutzeroberfläche die SQL-freie Exploration der Daten ermöglicht. Die Daten werden dabei in einer tabellarischen Struktur angezeigt, welche nach einzelnen Parametern geordnet oder durchsucht werden kann. Weiterhin besteht die Möglichkeit des Exports der Datenbank in eine *txt*[1]-Datei, wobei die Tabelleneinträge durch einen Zeilenumbruch und die Einträge innerhalb einer Zeile durch einen Tab, Komma oder Punkt separiert werden können.

[1] Eine *txt*-Datei oder auch *Plain Text File* ist ein Dokumententyp, dass das Speichern von Zeichen und Texten in unformatiertem Zustand ermöglicht [69].

Eine Aktion oder Vorgang eingetragen in das KIS „ICUData“ generiert einen bis mehrere Dateneinträge, die in zwei Relationen – der *obr* und der *obx* Relation/Tabelle – abgelegt werden. Diese Dateien umfassen für den Zeitraum von Juni 2006 bis August 2019 ca. 490.000 (in der obr) respektive 1.250.000 (in der obx) Zeilen bzw. Dateneinträge. Jeder Eintrag beinhaltet eine Reihe von weiteren Vorgangsinformationen, die in Unterabschnitt 6.1.1 eingehender beschrieben werden. Das Öffnen und Durchsuchen einer txt-Datei, die die Vorgänge dieses Zeitraums beinhaltet, dauert mehrere Minuten.

6.1.1 Dateneinträge des KIS „ICUData“ der Klinik für Neurologie des Universitätsklinikums Magdeburg

Die Struktur eines Dateneintrages in der Klinikdatenbank wird in Tabelle 6.1, Tabelle 6.2 und Tabelle 6.3 dargestellt.

Die obx Tabelle beinhaltet die Daten klinischer Vorgänge - im weiteren Verlauf auch Events genannt -, wobei jede Zeile der Datei ein zu einem bestimmten Zeitpunkt stattgefundenes Event darstellt. Ein Eintrag beinhaltet verschiedene Bezeichner/Identifier, beteiligte Personen, Art und Kategorie des Vorgangs, einen zugehörigen Wert und den Zeitpunkt - alle den in Tabelle 6.1 enthaltenen Attributen entsprechend.

Für den Prozess der Datenintegration sind folgende Attribute für die Einordnung eines Events besonders erwähnenswert: Jedes Event bekommt einen eigenen numerischen Bezeichner über das Attribut *Id* - folgend für ein einfacheres Verständnis auch *Event_id* genannt. Events werden Einträgen der obr Tabelle über die Attribut *Obr_id* und *Id_obr* zugeordnet, wobei die *Id_obr* der gleichen numerischen Folge wie die *Event_id* folgt. Die Art eines Events kann anhand des Attributs *Category* - einer groben Einteilung in ca. 12 Bereiche, die in Tabelle 6.4 aufgezählt werden - und der Gruppe der *Observation_X* Attribute - einer genauen textuellen Beschreibung des Events - festgestellt werden.

Ein klinischer Vorgang kann durch mehrere Events dargestellt werden, um z.B. neben einem numerischen Laborwert auch einen textuellen Kommentar zugehörig zu speichern. Ein solcher Zusammenhang in einem Set wird durch eine gleiche *Id_obr* verzeichnet. Für diese Events gibt es außerdem eine Set-interne Nummerierung, gespeichert in der *Set_id*. Das Attribut *Pat_id* beinhaltet die ID des dem Event zugehörigen Patienten. Das *Value* Attribut enthält den eingetragenen Wert eines Events, beispielsweise einen Blutdruckwert oder Untersuchungsergebnisse, Entlassungsbriefe usw. Das *Value_type_id* Attribut gibt dabei an, nach welchem Datentyp der Inhalt des *Value* Attributs interpretiert werden soll, wobei zu beachten ist, dass alle Werte in der txt Datei nur als String Da-

Attribut	**Wert**
Id	1000000001
Next_id	
Last_id	
Pat_id	100001
Obr_id	10000000000000000000000000000000001
Set_id	10001
Id_obr	8000000001
Category	1
Value_type_id	NM
Observation_id	MDD_NBP-mean
Observation_txt	MDD_NBP-mean
Observation_cs	DRV
Alternative_observation_id	NBP
Alternative_observation_txt	NBP
Alternative_observation_cs	DRV
Observation_subid	
Value	90
Unit_id	mmHg
Unit_txt	mmHg
Unit_cs	L
Reference_range	
Abnormal_flag_id	
Probability	
Abnormal_ref_id	
Result_status_id	F
Access_check	
Observation_ts	01.01.2000 12:00:00
Producer_id	
Producer_txt	
Producer_cs	L
Observer_uid	11111
Observer_gid	99
Insert_uid	11111
Insert_gid	99
Insert_time	01.01.2000 12:05:00
Hl7msg_id	
Obr_rowid	

Tabelle 6.1: Struktur eines Dateneintrages der *obx*-Relation. Die Ausprägungen der Werte sind Beispiele und frei erfunden.

Attribut	**Wert**
Id	8000000001
Next_id	
Last_id	
Obr_id	10000000000000000000000000000000001
Pat_id	123456
Quell_pat_id	123456
Name	Max
Vorname	Mustermann
Geb_dat	01.01.1980 00:00:00
Fall_id	1234567
Quell_fall_id	
Episode_id	12345678901234
Set_id	10000
Placer_id	1234567890
Placer_appl_id	
Filler_id	10000000000000000000000000000000001
Filler_appl_id	
Service_id	ICUFILES
Service_txt	ICUFILES
Service_cs	ID8
Observation_ts	01.01.2000 12:00:00
Observation_end_ts	01.01.2000 12:00:00
Collection_vol_qty	
Collection_vol_unit	
Collector_uid	
Collector_gid	
Specimen_action_id	
Danger_id	
Danger_txt	
Danger_cs	L
Clinical_info	IP^KNEU-62714.imed.uni-magdeburg.de^554 ^NEU14^NEU14-5^NEU14-08^(ORUZ02:ID)
Received_ts	
Source_code_id	NORM
Source_code_txt	NORM
Source_code_cs	L
Source_additives	
Source_text	
Source_text	
Source_site_txt	
Source_site_cs	
Source_site_cs	5
Source_smod_txt	5
Source_smod_cs	L

Tabelle 6.2: Teil 1 der Struktur eines Dateneintrages der *obr*-Relation. Die Ausprägungen der Werte sind Beispiele und frei erfunden.

Attribut	**Wert**
Order_provider_uid	
Order_provider_gid	
Op_tel_typ	
Op_tel_lkz	
Op_tel_vw	
Op_tel_nr	
Op_tel_dw	
Placer_field_1	
Placer_field_2	
Filler_field_1	
Filler_field_2	
Result_ts	
Charge_value	
Charge_code	
Service_section_id	
Result_status_id	F
Result_copies_uid	
Result_copies_gid	
Transport_mode_id	
Reason_id	
Reason_txt	
Reason_cs	L
Principal_uid	
Principal_gid	
Assistent_uid	
Assistent_gid	
Technician_uid	
Technician_gid	
Transcript_uid	
Transcript_gid	
Scheduled_ts	01.01.2000 12:00:00
Insert_uid	
Insert_gid	
Insert_time	01.01.2000 12:00:00
Hl7msg_id	
Op_tel_nr_2	
Op_tel_nr_3	
Orig_ipid	

Tabelle 6.3: Teil 2 der Struktur eines Dateneintrages der *obr*-Relation. Die Ausprägungen der Werte sind Beispiele und frei erfunden.

Category	**Kategorie**	**Inhalte**
0	Medikamente	
1	Vitalwerte	Blutdruck, Puls, Sauerstoffsättigung, ...
2	Vitalwerte	Atemfrequenz, Beatmungszustand, ...
3	Patientenpflege	Mobilität, Lagerung, Verbandswechsel, ...
4	Patientenpflege	Körperpflege, Prophylaxe, Hautzustand, ...
5	Ausscheidungen	Urin, Stuhlgang, ...
6	Laborwerte	
7	Untersuchungen	Neurostatus, Motorik, Schmerzempfinden, ...
8	Ärztliche Kommunikation	Visite, Arztbriefe, Entlassungen, ...
9	Administration	Formulare der Patientenaufnahme, Anamnese, Verlegungen, ...
10	Medikamente	
14	Skalen-Tests	Sturzrisiko-Skala, MRE-Screening, ...

Tabelle 6.4: Werte des Attribut *Category*. Es handelt sich hierbei um eigene Erkenntnisse aus der Untersuchung der obx Tabelle.

tentyp vorhanden sind. Aufschluss für den Zeitpunkt des Events gibt das *Observation_ts* Attribut, dass einen auf die Sekunde genauen Zeitstempel beinhaltet.

Die obr Tabelle beinhaltet Meta Informationen in Zusammenhang zu den Event Einträgen der obx Tabelle. Es werden Patienten- und IT-Infrastruktur-Informationen gespeichert, die zeigen, zu welcher Person, an welchem Terminal, zu welcher Zeit, von welchem Personal ein Event eingetragen wurde. Jeder Eintrag wird durch eine eigene Ausprägung des *Id* Attributs gekennzeichnet. Das *Id_obr* Attribut der obx Tabelle ist äquivalent zu dieser ID - folgend für ein einfacheres Verständnis auch in der obr Tabelle als *Id_obr* bezeichnet. Ein Eintrag kann in den Attributen *Pat_id, Name, Vorname* und *Geburtsdatum* persönliche Patienteninformationen speichern.

Der Informationsgehalt dieses Datensatz ist durch die Verbindung von Individuen zu bestimmten Vorgängen geprägt. Die verschiedenen IDs zeigen unmissverständliche Beziehungen der Daten auf. Im Laufe der Analyse der Daten stellte sich heraus, dass vergebene Patienten IDs - in obx und obr *Pat_id* - nicht eindeutig zuordnenbar sind. So gibt es auftretende Fälle in der obr Datei, in denen eine Patienten ID mehrere Male an unterschiedliche Personen/Patienten vergeben wurde. Außerdem kann es auch vorkommen, dass ein Patient mehrere verschiedene IDs bekommen hat. Dies hat zur Folge, dass Events aus der obx Datei nicht in allen Fällen eindeutig mit bestimmten Personen in Verbindung gesetzt werden können, da in der obx Tabelle das *Pat_id* Attribut das einzige Patienten-identifizierende Datum ist.

Die Struktur der Datenbank und der Datenrepräsentation folgt aus der Implementierung des KIS und den eigenen durchgeführten Modifikationen und Spezifikationen der Dateneingabe durch das Krankenhauspersonal selbst. Medizinische Standards der Datenrepräsentation (siehe Kapitel 3) werden bisher allerdings nicht strukturell verwendet [70]. Sie können jedoch als Ausprägung eines eigenen Events und dessen *Value* Attribut auftreten und so zumindest als punktuell vorhanden angesehen werden.

Verwendete Medikamente wurden ursprünglich basierend auf der *Roten Liste* Stand 1986-88 vermerkt und gespeichert. Die Rote Liste ist ein Verzeichnis für Medikamente und Arzneimittel [71]. Seitdem wurden Informationen bezüglich der Medikamente nach eigenem Ermessen geändert und neue Medikationen hinzugefügt, sodass nicht von einer standardisierten Medikamenten-Datenbasis ausgegangen werden kann. Seit 2019 werden LOINC Codes für die Repräsentation von Labordaten verwendet. Verwendete Kodierungen und Identifizierungen übriger medizinischer Daten gelten klinikintern oder KIS „ICUData“-intern.

6.2 Verwendete Daten der Arbeitsgruppe MedDigit

Die Forschungsgruppe „MedDigit - Medizin und Digitalisierung“ der Klinik für Neurologie des Universitätsklinikums Magdeburg benutzt für Forschungszwecke gesammelte patientenspezifische Daten. Für die verschiedenen Forschungsprojekte und Studien der Gruppe werden zur Zeit größtenteils Daten aus den medizinischen bildgebenden Verfahren verarbeitet und analysiert.

Die für die Forschungsarbeit der Gruppe zur Verfügung stehenden Daten sind aus dem KIS der Klinik für Neurologie entnommen. Eine systematische Verwendung oder Analyse der klinischen Datenbasis des Universitätsklinikums Magdeburg findet nicht statt.

6.3 Theoretische Anforderungen an eine Datenbank für klinische Routinedaten

Die Anforderungen entstehen in enger Zusammenarbeit mit der Forschungsgruppe „MedDigit - Medizin und Digitalisierung"der Klinik für Neurologie des Universitätsklinikums Magdeburg. Prinzipiell soll eine Verfügbarkeit von klinischen Routineinformationen hergestellt werden, die dem zukünftigen Forschungsbetrieb der Gruppe neue Möglichkeiten eröffnen kann. So können auch Bemühungen in den Bereichen stati-

stische Auswertungen und Deep Learning Modelle zur Hypothesengenerierung weiter gefördert werden.

Die klinischen Patientendaten enthalten demografische Daten, Patientenpflege- und Behandlungsdaten, Diagnosen und Arztbriefe, Medikationen, Laborwerte und stehen in einem Plain-text-Dateiformat verfügbar. Die klinischen Daten sollen auf die IT-Infrastruktur der Arbeitsgruppe MedDigit übertragen werden und eine Verfügbarkeit für die Forschungsarbeit hergestellt werden. Es soll eine relationale Datenbank erstellt werden, die diese Werte abbilden können soll. Folgende Anforderungen lassen sich formulieren:

R1 Die Daten sollen in einem passenden Datentyp abgebildet werden. Die rein textuell vorliegenden Daten sollen für eine vereinfachte Analyse und Verarbeitung in binäre Aussagen (Wahrheitswerte), bzw. Zahlendaten umgewandelt werden oder textuell verbleiben.

R2 Es sollen verschiedene Relationen für die ‚Themenbereiche' der Patientendaten erstellt werden, um Redundanzen zu verringern und eine Übersichtlichkeit des Datenbestandes herzustellen.

R3 Logischer und verständlicher Aufbau der Datenbank, um eine fortlaufende Nutzung zu gewähren ohne umfängliche Kenntnisse über Datenbanktechnologien besitzen zu müssen.

R4 Die spezifische Suche und Ausgabe von Daten soll ermöglicht werden.

R5 Die medizinischen Daten sollen anonymisiert werden können, um mit Nutzung und Veröffentlichung von Forschungsergebnissen keinen Patienten identifizierbar zu machen.

Mit der Umsetzung dieser Anforderungen kann der Einsatz von klinischen Daten der Universitätsklinik Magdeburg in der Arbeit der Forschungsgruppe MedDigit ermöglicht werden. Die Nutzung der Daten unterliegt dabei einer gewissen Lokalität, da diese bisher nur aus der „ICUData" KIS Struktur in eine überhaupt nutzbare Form gebracht worden wären. Wenn eine weitere Kompatibilität und Vergleichbarkeit der Daten über den eigenen Standort hinweg mit anderen Datenbanken der klinischen Domäne hergestellt werden soll, müssen weitere Schritte der Modellierung, Umstrukturierung und Übersetzung erfolgen und neue Anforderungen erfüllt werden.

Zwei Stufen der Kompatibilität können logisch definiert werden:

- *Strukturelle Kompatibilität*: Datenbanken teilen sich ein Datenmodell und besitzen damit einen gleichen Aufbau. Dies kann z.B. bedeuten, dass sie aus dem gleichen logischen Aufbau an Relationen bestehen.
- *Repräsentative Kompatibilität*: Datenbanken teilen sich eine gleiche Repräsentation ihres Inhalts. Für medizinische Daten könnte dieses z.B. bedeuten, dass alle diagnostischen Informationen in ICD-10-GM Codes (siehe Unterabschnitt 3.1.1) gespeichert werden.

Um eine nationale oder internationale Vergleichbarkeit oder Zusammenführung der Daten zu ermöglichen, sollten folgende zusätzliche Anforderungen berücksichtigt werden:

AR6 Die Datenbank soll nach einem frequent genutztem Datenmodell konzeptioniert sein, um eine *Strukturelle Kompatibilität* herzustellen. Ein Beispiel für ein Datenmodell aus der Domäne für medizinische Datenspeicherung ist das OMOP (siehe Unterabschnitt 3.1.3).

AR7 Die gespeicherten Daten sollen in einer frequent verwendeten Repräsentationsform vorliegen, um eine *Repräsentative Kompatibilität* herzustellen. Beispiele werden in Unterabschnitt 3.1.1 und 3.1.2 genannt.

AR8 Der Datenschutz und andere Rechte der Patienten sollen in allen Schritten der Veröffentlichung oder Teilung ihrer Daten national wie international gewahrt bleiben.

6.4 Analyse der theoretischen Anforderungen und Schlussfolgerung

Eine Erfüllung der in Kapitel 6.3 genannten Anforderungen würde eine Datenbank erstellen, die die Forschungsarbeiten der Arbeitsgruppe MedDigit maßgeblich erweitert. Die Analyse und Verwendung der im klinischen Alltag der Patientenversorgung entstehenden Daten des Universitätsklinikum Magdeburg bringt innovative Möglichkeiten für die Gewinnung neuer medizinischer Erkenntnisse. Durch eine Erschließung eines solchen Datenbestandes würde die Menge, der für die Wissenschaft zur Verfügung stehenden Daten, vervielfacht werden.

Basierend auf der Analyse vergleichbarer Vorhaben in Kapitel 5 müssen jedoch die gestellten Anforderungen reevaluiert werden. Das von der deutschen Bundesregierung geförderte und finanzierte MIRACUM-Projekt befindet sich nach 3 Jahren intensiver Ar-

beit noch in einer konzeptionellen und prototypischen Phase der Erstellung einer funktionsfähigen vernetzten Datenbank für medizinische Daten, die mit ihren Anforderungen grundsätzlich Ähnlichen folgt, wie in diesem Kapitel aufgestellt. Des Weiteren gibt es bisher keine spezifischen Vorgehensweisen einer Datenintegration der Datenbank eines KIS, wie der des „ICUData“ in Magdeburg. Bis zu ihrer Fertigstellung sind Konzeptionierungen und Implementierung darauf aufbauender Funktionalitäten wie die Anonymisierung identifizierender Daten, das Management von Patienteneinwilligungen, ihre Datennutzung, die Vernetzung und Planung standardisierter Modelle hinfällig. Das in dem MAGNET-Projekt verwendete Datenintegrationsprogramm beinhaltet einen mehrjährigen Entwicklungsaufwand und ist darüber hinaus nur mit den in Australien verwendeten IT-Systemen kompatibel. Bei einem gleichen Programm für das KIS „ICUData“ ist somit bis zu der Realisierung von einer erheblichen Zeitspanne auszugehen.

Aufgrund begrenzter zeitlicher Ressourcen im Rahmen dieser Bachelorarbeit ist es nicht möglich die Arbeit langjähriger Forschungsprojekte nationaler Konsortien nachzubilden und für die Erstellung einer lauffähigen Datenbank aufzuwenden. Um neue Erkenntnisse in dem Bereich des Managements klinischer Routinedaten zu erhalten, muss sich zunächst auf eine Analyse und Konzeption einer konkreten Strategie für die Datenintegration der KIS Datenbank der Klinik für Neurologie des Klinikums Magdeburg in eine neue, für die Forschung nutzbare, Datenbank konzentriert werden.

In den folgenden Kapiteln der Bachelorarbeit wird sich deswegen auf die grundlegenden Kernfunktionen und Anforderungen einer Datenintegration der „ICUData“ Datenbank in eine neu entworfene Datenbank mit den Anforderungen **R1** bis **R5** beschränkt. Die erweiternden Anforderungen **AR6** bis **AR8** für die Herstellung einer Kompatibilität mit anderen medizinischen Datenquellen sind Funktionalitäten, die in weiterführenden Arbeiten der Forschungsdatenbank behandelt werden sollten.

7 Konzept der Applikation

Die in den vorherigen Kapiteln betrachteten Aspekte der Datenbanktheorie, Beschaffenheit medizinischer Daten, der Wichtigkeit von Privatsphäre und Erkenntnisse aus verwandten Arbeiten fließen in einer Konzeption eines Prozesses der Datenverarbeitung nach den Anforderungen der Arbeitsgruppe zusammen. Zunächst wird eine generalisierte Pipeline projiziert, die den theoretischen Verlauf der Daten und Vorgänge aufzeigt. Folgend werden die einzelnen Stationen des Prozesses genauer beschrieben.

7.1 Die Pipeline für klinische Routinedaten

Eine Analyse und Verarbeitung von medizinischen Daten aus dem „ICUData“ System der Klinik für Neurologie der Universitätsklinik Magdeburg könnte auf der vorhandenen relationalen Oracle Datenbank des KIS stattfinden. Dies ist jedoch aus verschiedenen Gründen nicht empfehlenswert:

Ein direkter Zugang zu der Datenbank benötigt eine Einbindung in die Hardware- und Software-Struktur der Klinik. Der Aufbau der vorhandene IT-Infrastruktur und ihrer Funktionalität resultiert aus Anforderungen an die lokale Vernetzung, Langlebigkeit und Datenschutz sowie dem Bedürfnis einer digitalen Verwaltung von Informationen auf individueller Patientenbasis. Ein solches System ist nicht für die Nutzung in einer großflächigen Datenanalyse in Zusammenhang eines Forschungskontextes ausgelegt. Des Weiteren würde ein Anbindung eine Kompatibilität aller beteiligten Systeme auf technischer Ebene voraussetzen, damit die Stabilität und Funktionsfähigkeit gewährleistet werden kann. Da die KIS Datenbank eine Reihe an identifizierenden Informationen beinhaltet, ist eine Veröffentlichung der Daten im Rahmen von Forschungsarbeiten nicht möglich. Eine effektive Forschungsarbeit kann so nicht stattfinden.

Aus diesen Gründen wird eine Struktur aus getrennten Systemen und einer eigenen relationalen Datenbank für für Forschungszwecke verwendete Daten vorgeschlagen. Es

wird eine bewusste Trennung zu der kritischen IT-Infrastruktur der Klinik vorgenommen, um einen *Safe Space* der Datennutzung zu erzeugen. Eine Abschottung verhindert jegliche Einflussnahme der Arbeit der Forschungsgruppe auf das elektronische Dokumenationssystem des ärztlichen und pflegenden Personals. Eine separate Datenbank erlaubt eine Prozessierung und Verarbeitung der Daten im Rahmen eines experimentellen Umgangs ohne Sorge einer technischen Störung des Betriebes der Klinik oder Datenverlustes der Patientendaten.

Die Export-Funktion der Oracle Datenbank des „ICUData“ Systems der Klinik erlaubt eine portable Kopie des Datensatzes zu erstellen. Die dabei erzeugten txt Dateien der obx und obr Tabellen können in das System der Forschungsgruppe übertragen werden und befinden sich damit ab diesem Zeitpunkt außerhalb der IT-Infrastruktur der Patientenversorgung. Innerhalb des Systems der Forschungsgruppe kann eine weitere Verarbeitung der erhaltenen Dateien der klinischen Daten erfolgen, um die Anforderungen aus Kapitel 6 zu erfüllen.

Für eine effiziente und langfristig Nutzung der Daten ist der Einsatz einer eigenen relationalen Datenbank, aufgrund der in Kapitel 2 besprochenen Funktionen und Vorteile, den Anforderungen entsprechend sinnvoll. Eine relationale Datenbank in einer, wie bisher beschriebenen eigenverantwortlichen Infrastruktur gibt eine Freiheit der Datennutzung und Flexibilität in der Nutzung weiterer technischer Systeme und Programme, die mit einer solchen Datenbank interagieren können.

Der Schritt der Überführung der Rohdaten der Tab-separierten txt-Dateien in die Datenbank findet in der Datenintegration statt. Damit werden die in reinem Freitext vorhandenen Daten nach Feldern und Attributen selektiert, nach Datentyp kategorisiert und durch Einsatz von SQL-Querys in eine Datenbank eingefügt. Es findet eine umfangreiche Anonymisierung und Pseudonymisierung der erhaltenen medizinischen Daten statt. Identifizierende Informationen, wie Namen, Geburtstage, Patienten- und Event-IDs sowie ärztliche Kommunikationen werden gefiltert und durch ein kontinuierliches Pseudonym ersetzt, welches eine Vernetzung medizinischer Vorgänge an ein nun anonymes Individuum erlaubt. Um eine Nach- und Rückverfolgung dieses Prozesses zu sichern, wird eine weitere Datenbank erstellt, die die gefilterten identifizierten Informationen – die IDAT oder Identifying Data – enthält und ein dazugehöriges Pseudonym. So können auch im Falle wichtiger neuer Erkenntnisse, die die Gesundheit des Patienten betreffen, die wahre Identität wiederhergestellt und entsprechende Schritte eingeleitet werden, um der Person diese Informationen zukommen zu lassen. Der Verschluss der Identität und Wahrung der Privatsphäre von Patienten erlaubt eine freie Forschungs-

tätigkeit. Somit sind die auf der Datenbank des Forschungsteams liegenden Daten frei von identifizierenden Informationen, frei von technischen Abhängigkeiten zu anderen Systemen des Krankenhauses und frei im Umgang des forschenden Personals.

Die in Unterabschnitt 6.1.1 erwähnte Unsicherheit in der Zuordnung von Events zu einzelnen Patienten wird durch eine Dokumentation der Beziehungen verschiedener Patienten und ihrer IDs verringert. Unsicherheiten wie, die Zuordnung mehrerer IDs zu einem Patienten und umgekehrt mehrerer Patienten zu einer ID, werden also schon vor bzw. während des Pseudonymisierungs- und Anonymisierungsprozesses überprüft, erkannt und ohne identifizierende Elemente in der Forschungsdatenbank vermerkt.

Die Datenintegration findet auf einem separaten System der Forschungsgruppe statt. Die Isolation dieses Verarbeitungsprozesses der exportierten txt Dateien auf ein einzelnes System reduziert den Umgang mit sensitiven persönlichen Informationen auf ein Mindestmaß und ermöglicht es, Zugriffsbeschränkungen aufzuerlegen, die nur Abrufe von autorisiertem Personal zulassen. Die Datenbank der medizinischen Daten – die MDAT oder Medical Data - befindet sich auf einem eigenen System, auf das die Mitarbeiter der Forschungsgruppe aufgrund der vorherigen Verarbeitung freien Zugang haben.

In Abbildung 7.1 wird der Prozess dieser Datenverarbeitung dargestellt.

7.1.1 Station 1: Export

Das KIS der Klinik für Neurologie beinhaltet ein Zugriffsprogramm für die Datenbank des „ICUData“ Systems. Integriert in dieses Zugriffsprogramm ist eine Export-Funktionalität, die den Inhalt der relationalen Oracle Datenbank in eine Tab-separierte Plain Text File bzw. txt Datei, wie in Unterabschnitt 6.1.1 besprochen, überträgt. Die Texteinträge der einzelnen Attribute der Datenbank sind dabei stets durch ein Tab Symbol getrennt. Die Tupel der Datenbank werden durch einen Zeilenumbruch getrennt, sodass jeder Tupel in einer neuen Zeile dargestellt wird. Die Export-Funktion erlaubt weiterhin einen Zeitraum der zu exportierenden Daten auszuwählen und nur Daten innerhalb zu exportieren.

Die Ausführung des Exportes wird manuell durch eine autorisierte Person der Klinik für Neurologie durchgeführt. Die entstandenen txt Dateien werden danach an die Arbeitsgruppe MedDigit übergeben, die die Dateien manuell auf das System der Station 2 überträgt.

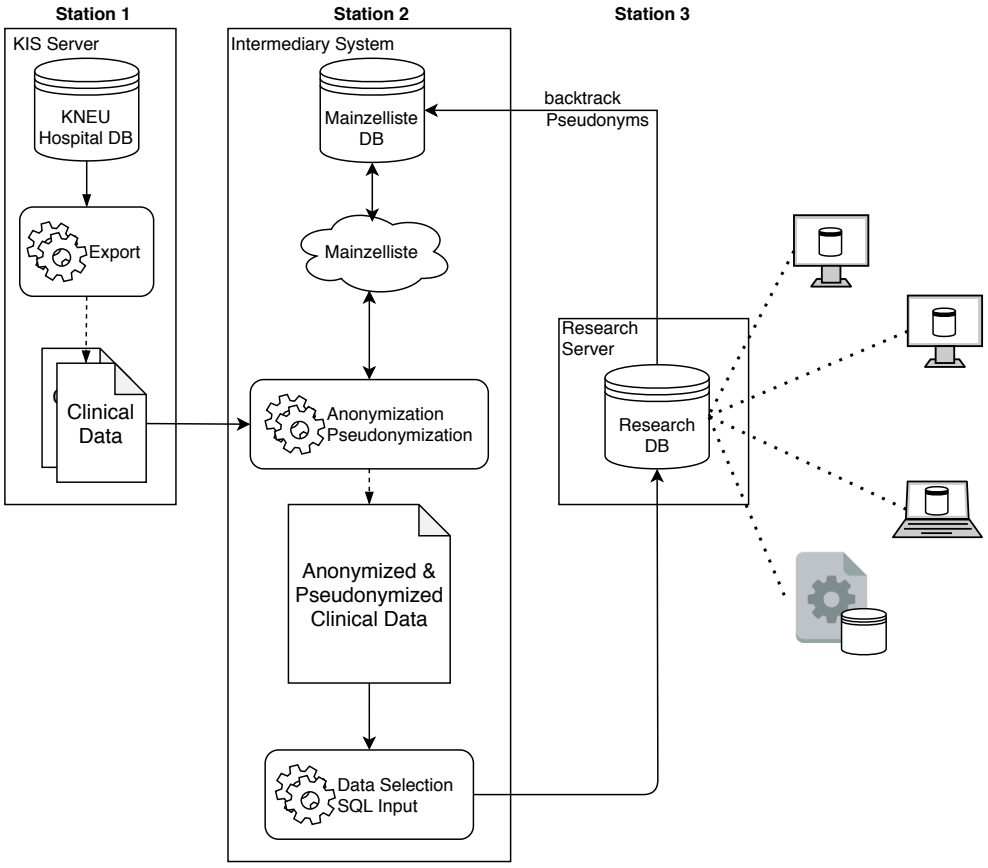

Abbildung 7.1: Konzeption der Pipeline, welche die Daten der Klinik für Neurologie durchlaufen, um für den Forschungsbetrieb nutzbar zu werden.

7.1.2 Station 2: Intermediary System und Datenintegration

Die Tab-separierten txt Dateien der obr und obx Tabelle befinden sich auf einem System der Forschungsgruppe (siehe Station 2 in Abbildung 7.1).

Auf diesem System befindet sich die Funktionalität der Datenintegration. Hierbei werden die textuellen Informationen automatisiert aus der Struktur der txt Dateien ausgelesen und eine Bewertung des Inhalts wird durchgeführt. Das Schema der Datenbank des Forschungsteams bestimmt welche Daten entnommen werden sollen. Bevor eine Übertragung der ausgelesenen Daten auf die Datenbank stattfindet, werden die Daten anonymisiert und pseudonymisiert. Das „ICUData“ Programm ermöglicht viele Vorgänge der Patientenversorgung in einer Freitextform festzuhalten. Dies können beispielsweise Arzt- und Entlassungsbriefe sein, die verschiedene identifizierende Informationen wie Namen, Geburtsdaten und Adressen beinhalten können. Eine Analyse und sichere

Selektion von Freitexten, die keine solche Informationen enthalten ist eine technisch komplexe Anforderung, dessen Umsetzung den Umfang dieser Arbeit übersteigen würde. Unter diesem Aspekt ist eine Filterung dieser Freitext-Events und Nicht-Aufnahme in die Datenbank eine Methode, um zu einer gesicherten Anonymisierung zu gelangen. Außerhalb der Freitexte gespeicherte identifizierende Informationen werden durch eine Pseudonymisierung entzogen und ersetzt. Die einzige Ausnahme bildet das Geburtsdatum. Es ist eine Information mit einem hohen Grad an persönlicher Identifizierbarkeit und sollte aus diesem Grund nicht in den Daten der Forschungsdatenbank auftreten, jedoch ist das Alter eines Individuums höchst wichtig für die Einordnung klinischer Vorgänge und Prozeduren. Gelöst wird dieser Zwiespalt durch die Ersetzung von Geburtsdaten mit einer direkten Berechnung des Alters in Jahren im Verhältnis zum Zeitpunkt der Berechnung. Des Weiteren wird zu jedem klinischen Event das Alter des Patienten zum Zeitpunkt des Stattfindens des Events berechnet, sodass eine Wahrung des Datums ohne Informationsverlust vorgenommen werden kann. Für die Berechnung einzigartiger Pseudonyme für verschiedene Patienten, um die wichtige Verbindungen der klinischen Events zu Individuen zu gewährleisten, bedarf es eines Algorithmus. Abschließend werden IDs in Zusammenhang der Events dem gleichen Ablauf unterzogen und durch Pseudonyme ersetzt.

Um diesen Prozess eine Bilateralität zu gewähren wird eine Datenbank eingerichtet, die sowohl die Eingaben des Pseudonymisierungsalgorithmus – die IDAT – als auch die Ausgaben – die Pseudonyme – beinhaltet und in Relation setzt. Die Datenbank der identifizierenden Information befindet sich auf dem gleichen System, auf dem der Datenintegrationsprozess abläuft und ist nur innersystemisch zugänglich. Das in Unterabschnitt 4.1.1 beschriebene Pseudonymisierungsprogramm Mainzelliste kann für die Erstellung der Pseudonyme und die Verwaltung einer Datenbank für identifizierende Informationen eingesetzt werden.

Für jedes ausgelesene Event werden automatisiert SQL-Anfragen an die auf Station 3 liegende Forschungsdatenbank erstellt und über das lokale Netzwerk der Arbeitsgruppe MedDigit versendet. Die Anfragen beinhalten die Direktive einer Dateneintragung des klinischen Vorgangs des einzutragenden Events. Die mitversendeten Informationen sind dabei jene, die aus den txt Dateien dekodiert wurden und in die Attribute des Forschungsdatenbank passen; ausgenommen sind, die zu filternden Events und identifizierenden Informationen.

7.1.3 Station 3: Datenbank

Die dritte Station stellt das System für die freie Nutzung für Forschungszwecke dar. Hier wird eine relationale Datenbank für die klinischen Routinedaten aus der Klinik für Neurologie erstellt, die über das lokale Netzwerk der Arbeitsgruppe ansteuerbar ist. Über die SQL-Sprache können Anfragen zur Datenselektion an die Datenbank gestellt werden als der Startpunkt für weitere Untersuchungen und Forschungsarbeiten.

Die Struktur der Datenbank, das Datenbankschema, folgt der der bestehenden obx und obr Tabellen. Für eine simple Translation der Daten in die technische Struktur der Forschungsgruppe werden Attributnamen direkt übernommen und die Beziehungen der Attribute bleiben bestehen. Bisher angeeignetes Wissen über den Datenbestand der Klinik innerhalb und außerhalb der Forschungsgruppe kann so problemlos auf die neue Datenbank übertragen und angewandt werden.

Die Struktur wird in Abbildung 7.2 dargestellt. Es werden 5 Tabellen definiert, dessen Attribute und Werte aus den obr und obx Tabellen entnommen werden:

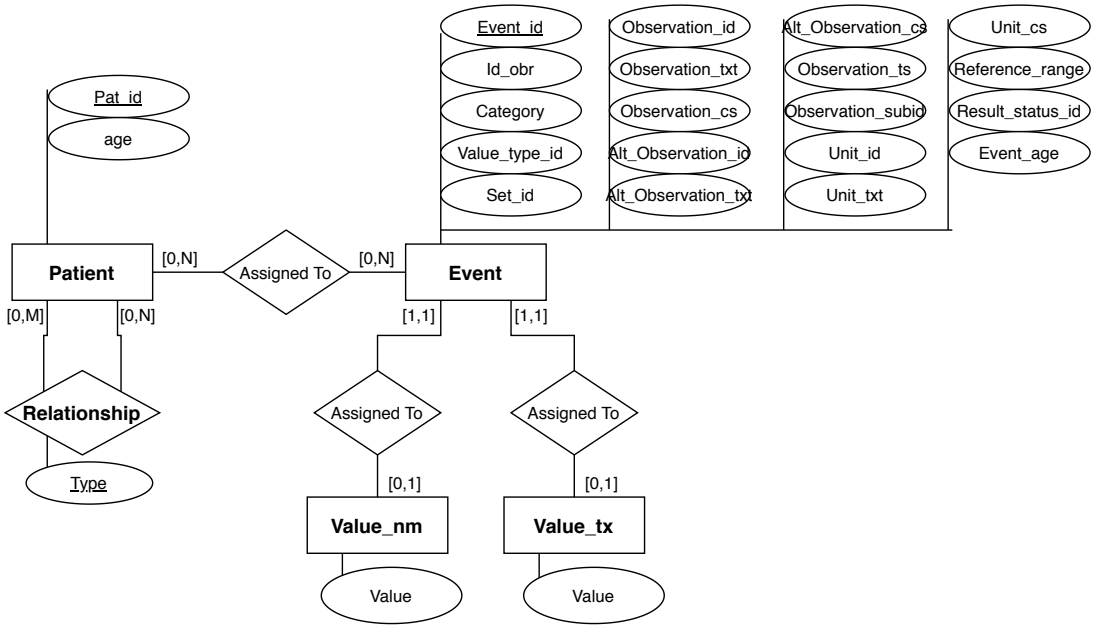

Abbildung 7.2: Konzeption eines Entity-Relationship-Diagramms der Forschungsdatenbank.

Patient Die Tabelle beinhaltet das Attribut *Pat_id* und *age*. Und wird mit den Daten der obr Tabelle befüllt. Das *Pat_id* Attribut beinhaltet hierbei das erzeugt Pseudonym eines Patienten. Das Attribut *age* beinhaltet das aktuelle Alter eines Patienten in Jahren.

Relationship Die Tabelle beinhaltet die Beziehungen von Patienten bzw. ihrer IDs. So gibt es eine Zuordnung eines Anker-Pseudonyms - dem *anchor_pseu* Attribut - zu einer anderen Pseudonym - dem *pseu* Attribut - und einer Beziehungstyp - dem *type* Attribut. Handelt es sich vermutlich um den gleichen Patienten ist die Ausprägung des Typ 'S' (Same); sind es verschiedene Patienten mit der gleichen ID, ist die Ausprägung des Typ 'D' (Different).

Event Die Tabelle enthält den Großteil der medizinischen Informationen der klinischen Vorgänge aus der obx Tabelle. Die Attribute *Event_id, Id_obr* und *Pat_id* enthalten ihre Pseudonym Pendant. Das *Event_age* beinhaltet das Alter des Patienten in Jahren zu dem Zeitpunkt des Events. Die restlichen übernommenen Attribute können Abbildung 7.2 entnommen werden.

Value_nm Die Tabelle enthält das Attribut *Event_id*, das das Pseudonym einer Event-ID enthält und so auf ein bestimmtes Event in der Event Tabelle verweist. Das zweite Attribut *Value* beinhaltet den eigentlich gespeicherten Wert in Zusammenhang eines klinischen Events. Das NM im Namen der Tabelle weist darauf hin, dass in dieser Tabelle nur Werte von Events beinhaltet, denen numerische Werte zugehören.

Value_tx Die Tabelle ist identisch mit der Value_nm Tabelle und beinhaltet im *Value* Attribut textuelle Werte die Events zugeordnet sind.

8 Implementierung

Die Implementierung der in Kapitel 7 konzeptionierten Datenpipeline bedarf einer Initialisierung und Nutzung verschiedener Systeme und Technologien. Die Station 2 ist eine virtuelle Maschine (VM) mit dem Ubuntu Linux Betriebssystem, die auf dem Hauptserver mit dem Fedora Linux Betriebssystem der Arbeitsgruppe läuft. Alle Prozesse der Datenintegration finden auf dieser VM statt. So ist eine abgeschlossene Umgebung geschaffen für die Verarbeitung sensibler und persönlicher Informationen. Die Station 3 ist die relationale Datenbank für die Forschungsdaten. Diese läuft auf dem Hauptserver der Arbeitsgruppe.

8.1 Mainzelliste

Für den konzeptionierten Anonymisierungs- und Pseudonymisierungsvorgang wird das in Kapitel 4 Beschriebene Mainzelliste-Programm verwendet. Es steht eine Docker Installation für eine komplett funktionsfähige Mainzelliste zur Verfügung, die sowohl das Mainzelliste Programm, den Hosting Service als auch eine PostgreSQL Datenbank enthält. Somit müssen keine weiteren Programme oder Dienste zusätzlich für einen Betrieb der Mainzelliste installiert werden.

Es werden zwei Instanzen des Mainzelliste-Programms initialisiert, um einen getrennten Pseudonymisierungsvorgang für Patienten, ihren persönlichen Informationen und IDs, und den IDs der Events zu erhalten. Jede Instanz wird eigenständig über Docker betrieben, hat eine eigene Konfiguration und eine eigene angeschlossene Datenbank, die die sensiblen (zu pseudonymisierenden) Daten beinhaltet.

Patient Die Mainzelliste *Patient* beinhaltet die Attribute Vorname, Nachname, Geburtsdatum und die Patienten ID als eine externe ID. Eine erfolgreiche Eingabe dieser Daten erzeugt ein einzigartiges Pseudonym für dieses Individuum. Sollten die

IDAT schon einem anderen Eintrag mit einer anderen Patienten ID zugeordnet sein, werden Vor- und Nachname mit einem '_' Symbol erweitert und erneut in die Mainzelliste eingetragen, um für den Patienten auch unter der weiteren Patienten ID ein eigenes Pseudonym zu erzeugen. In diesem Fall wird auch eine Eintragung in der Relationship-Tabelle der Forschungsdatenbank vorgenommen, mit dem Pseudonym des schon vorhandenen Eintrags als Anker, dem Pseudonym des neuen Eintrags und dem Typ 'S'. Sollte eine Patienten ID in der Mainzelliste schon eingetragen sein, aber nun ein anderer Patient mit anderen IDAT neu eingetragen werden, wird die Patienten ID mit einem '_' Symbol erweitert und der Patient wird erneut in die Mainzelliste eingetragen. So werden für zwei Patienten mit verschiedenen IDAT zwei verschiedene Pseudonyme erzeugt. Auch in diesem Fall gibt es eine Eintragung in die Relationship-Tabelle mit den beiden Pseudonymen der beteiligten Patienten und dem Typ 'D'.

Über das Pseudonym oder die Patienten ID können die eingetragenen Daten der Attribute wieder abgerufen werden.

Event Die Mainzelliste *Event* beinhaltet keine Attribute, sondern nur die Event ID als externe ID. Somit besteht eine Konvertierung einer Event ID zu einem einzigartigen Pseudonym.

Die entstehenden Pseudonyme beider Mainzelliste Instanzen können und werden sich doppeln, sind aber für beide Bereiche individuell. Somit ist, wenn der Bereich des Pseudonyms bekannt ist, der referenzierte Dateneintrag der jeweiligen Mainzelliste Instanz eindeutig.

8.2 Datenbank

Als relationale Datenbank wird eine PostgreSQL Datenbank auf dem Linux Server der Arbeitsgruppe installiert und initialisiert. PostgreSQL ist ein Open-Source objektrelationale Datenbankmanagementsystem, das das Prinzip einer klassischen relationalen Datenbank mit komplexen Datentypen (Enums, Arrays, Listen, usw.) ergänzt und eigene Definitionen von Datentypen und Funktionen erlaubt [72]. Die Forschungsdatenbank benutzt derzeit nur die relationalen Funktionalitäten, aber der Einsatz einer PostgreSQL Datenbank erlaubt zukünftige Erweiterungen der Funktionlität der Datenbank.

Der Zugriff auf die Datenbank kann und wird in der Konfiguration auf bestimmte IP-Adressen beschränkt. Die Interaktion mit der PostgreSQL Datenbank erfolgt auf dem

Attribut	Datentyp	Kondition
Pat_id	character varying	Primary key
age	integer	

Tabelle 8.1: Das Relationsschema der Patient Tabelle der Forschungsdatenbank.

Attribut	Datentyp	Kondition
anchor_pseu	character varying	Primary key
pseu	character varying	Primary key
type	character varying	Primary key

Tabelle 8.2: Das Relationsschema der Relationship Tabelle der Forschungsdatenbank.

Host System über das Kommandozeilen-Programm *psql*. Für eine einfachere Interaktion auf externen Systemen wird das *pgAdmin4* Programm genutzt, das eine Wartung der Datenbank über eine grafische Benutzeroberfläche erlaubt.

Die in Kapitel 7 konzeptionierte Datenbank resultiert in Tabelle 8.1, 8.2, 8.3, 8.4 und 8.5 dargestellten Relationen, die über entsprechende SQL-Querys erzeugt werden.

8.3 Datenverarbeitung und -integration

Die Datenverarbeitung verläuft automatisiert über ein im Rahmen dieser Arbeit verfasstes Programm bzw. Skript. Das Skript wurde in der Python Sprache programmiert und vereint alle Schritte der Datenintegration. Die Tab-separierten txt-Dateien werden in ein Matrix-Dataframe eingelesen, wobei Zeilen übersprungen werden, deren Anzahl an Tab Symbolen die Anzahl an möglichen Attributen übersteigt und somit strukturell uninteressant für eine weitere Verarbeitung sind.

Es folgt zunächst die Verarbeitung der identifizierenden Patienteninformationen der obr Datei. Relevant und automatisiert verarbeitbar sind die als Attribute vorhandenen Merkmale Name, Vorname, Geburtsdatum und Patienten-ID. Das Dataframe der eingelesenen obr Datei wird nun zeilenweise verarbeitet. Zeilen, die gültige Eintragungen in den genannten Attributen besitzen, werden gewählt und ihre Daten werden über *addPatient* Anfrage an die Mainzelliste Patient geschickt. Die IDAT werden in der Datenbank der Mainzelliste gespeichert und es wird ein einzigartiges Pseudonym für diesen Patienten erzeugt. Aus dem Geburtsdatum wird das aktuelle Alter berechnet. Das Pseudonym und aktuelle Alter werden nun über eine SQL-Query in die Tabelle Patient der Datenbank eingetragen. Dieser Ablauf wird für alle Zeilen, der obr Datei vollzogen.

Attribut	Datentyp	Kondition
Event_id	character varying	Primary key
Pat_id	character varying	Foreign key
Id_obr	character varying	
Category	integer	
Set_id	character varying	
Value_type_id	character	
Observation_id	character varying	
Observation_txt	character varying	
Observation_cs	character varying	
Alternative_observation_id	character varying	
Alternative_observation_txt	character varying	
Alternative_observation_cs	character varying	
Observation_ts	timestamp	
Observation_subid	character varying	
Unit_id	character varying	
Unit_txt	character varying	
Unit_cs	character varying	
Reference_range	character varying	
Result_status_id	character varying	
Event_age	integer	

Tabelle 8.3: Das Relationsschema der Event Tabelle der Forschungsdatenbank.

Attribut	Datentyp	Kondition
Event_id	character varying	Primary key Foreign key
Value	double	

Tabelle 8.4: Das Relationsschema der Value_nm Tabelle der Forschungsdatenbank.

Attribut	Datentyp	Kondition
Event_id	character varying	Primary key Foreign key
Value	character varying	

Tabelle 8.5: Das Relationsschema der Value_tx Tabelle der Forschungsdatenbank.

Anschließend wird die Integration der klinischen Vorgänge – der Events – der obx Datei vorgenommen. Um eine komplette Anonymisierung der Daten sicherzustellen, wird eine Reihe an Filterungen und Pseudonymisierungen vorgenommen. Das Skript beinhaltet eine frei konfigurierbare Filter- und Allowlist: die Listen beinhalten Attribut-Wert Paare, die eine bestimmte Ausprägung in den obx Daten respektive verbieten und filtern oder explizit erlauben. Das Dataframe der eingelesenen obx Datei wird nun ebenfalls zeilenweise verarbeitet. Das Ergebnis eines Vergleiches der Daten mit den Filtern gibt an, ob die Zeile verworfen oder weiterverarbeitet wird. Über diesen Filter können zunächst kategorisch Events ausgeschlossen werden, die viele identifizierende Informationen als Freitexte in ihren Werten speichern und so nicht automatisch erkennbar sind. Soll eine Zeile verwendet also in der Datenbank gespeichert werden, müssen drei IDs pseudonymisiert werden: die Patienten ID, die Event ID und die Obr ID. Die Patienten ID entspricht dabei derjenigen, die in der obr Datei vorhanden ist, und bereits mit weiteren persönlichen Informationen in die Mainzelliste eingefügt wurde. Mit einer Anfrage der Patienten ID erhält man das bereits erzeugte zugehörige Pseudonym. Die Event und Obr ID werden mit Anfragen an die zweite Mainzelliste Instanz pseudonymisiert. Eine so anonymisierte und pseudonymisierte Zeile der obx Datei kann nun über eine SQL-Query in die Tabelle Event der Datenbank eingetragen werden. Der gespeicherte Datenwert eines Events in dem Attribut *Value* wird je nach Datentyp in die Tabelle Value_nm, bei einem numerischen Wert oder in die Tabelle Value_tx, bei einem textuellen Wert, eingefügt.

9 Evaluation und Diskussion der Ergebnisse

In diesem Kapitel wird der Prozess der Datenintegration und die für die Forschung entwickelte Datenbank evaluiert und bewertet. Dafür werden die in Kapitel 6 gestellten Anforderungen herangezogen und mit den erreichten Ergebnissen verglichen. Zunächst wurden alle Aspekte der in Kapitel 7 konzeptionierten Datenpipeline implementiert, installiert und erfolgreich ausgeführt. Die Evaluierung ist in drei Bereiche aufgeteilt, die jeweils für sich betrachtet und analysiert werden, welche Anforderungen sie erfüllen.

9.1 Skript und Datenverarbeitung

Zwei Mainzelliste Instanzen laufen erfolgreich als Docker Programme auf der virtuellen Linux Maschine, die auf dem Server der Klinik läuft. Das Datenverarbeitungsskript wurde erfolgreich als Python Programm ausgeführt. Die Verarbeitung der exportierten obr und obx Dateien durch das erstellte Skript verläuft ohne funktionsstörende Fehler. Die zur Verfügung stehenden obr und obx Dateien mit 493.732 und 1.256.532 Zeilen bzw. Einträgen werden erfolgreich eingelesen.

Eine große Qualität der erstellten Software ist, dass ein wiederholtes Einlesen problemlos möglich ist. Dies bedeutet, dass die Initialisierung der Datenbank und der Pseudonymisierung durch die Mainzelliste nicht nur für eine einmalige Verarbeitung der Dateien verwendbar ist, sondern eine fortlaufende Erweiterung des Datensatzes möglich ist. So führt eine erneute Ausführung des Skriptes mit neuen Dateneinträgen zu einer Erweiterung des Datenbestandes der Forschungsdatenbank. Auch Dateneinträge, die schon in der Datenbank vorhanden sind, werden korrekt verarbeitet. Diese Flexibilität erlaubt eine Verarbeitung aller Dateneinträge unabhängig der gewählten Zeiträume und Überschneidungen mit bereits verarbeiteten Daten.

Das Skript bietet des Weiteren eine Reihe an einfach konfigurierbaren Parametern, die bestimmen, welche Attribute aus welchen Dateien ausgelesen werden. Erweiterbare

Filter- und Allow-Listen geben eine Möglichkeit, von ganzen Kategorien an Daten bis hin zu einzelnen Events zu filtern. So können Events für die Verarbeitung gefiltert werden, die sensible, Personen-identifizierbare Informationen enthalten.

Die gewählte Pseudonymisierungs-Strategie für die IDs der Events und Patienten ist funktionsfähig und ersetzt erfolgreich und kontinuierlich sensible Daten durch abstrakte Pseudonyme, die eine Nach- und Weiterverfolgung möglich machen. Funktional problematisch ist die fundamentale Tatsache der Quelldaten, dass Patienten IDs zu mehreren verschiedenen Patienten zugewiesen werden können. Dies erhöht die Komplexität der Verarbeitung und Pseudonymisierung um ein Vielfaches. Die in Kapitel 8 besprochene Lösung ist funktionell, aber nicht allumfassend. Ein nicht gelöstes Problem stellt ein späteres Nachtragen von IDAT zu einer bestehenden Patienten ID dar: Es kann nicht überprüft werden, ob die Ergänzungen mit schon vorhandenen Patienteneinträgen der Mainzelliste übereinstimmen. So können Beziehungen, wie sie in der Tabelle Relationship der Forschungsdaten festgehalten werden, nicht nachträglich erkannt und dokumentiert werden. Das Mainzelliste-Programm ist nicht für diese Fälle von mehreren Patienten mit gleicher ID und einem Patienten mit mehreren IDs in seiner grundlegenden Funktionsweise ausgelegt. Die derzeitige Implementierung der Pseudonymisierung umgeht dies durch Manipulation der IDAT oder ID, kann aber keine nachträglichen Einträge von IDAT umsetzen. Unter den gegebenen Umständen ist diese Problematik nur zu lösen, wenn der Quelldatensatz der obr Datei durch ein anderes Patientenmanagament-System oder das Pseudonymisierungsprogramm Mainzelliste durch ein, für diese Fälle vorgesehenes, Programm ersetzt wird.

Der Prozess der Datenverarbeitung und -integration durch das erstellte Skript erfüllt die gestellten Anforderungen **R1** und **R5**. Zusätzlich wurde mit der Erstellung zahlreicher Konfigurationsparamter ein flexibler Umgang mit der Funktionalität der Datenverarbeitung gewährt, der auch eine zukünftige Addition neuer Daten für den Benutzer sehr unkompliziert gestaltet. Dies ist die wichtigste Ergebnis dieser Arbeit: es wurde ein essentieller erster Schritt der Datenintegration klinischer Routinedaten getan, der auch unabhängig der gewählten Struktur der Forschungsdatenbank neue Erkenntnisse über den Datenbestand der Klinik für Neurologie des Universitätsklinik Magdeburg liefert.

9.2 Datenbank und Datenbankstruktur

Die relationale PostgreSQL Datenbank wurde erfolgreich auf dem Server der Arbeitsgruppe installiert, konfiguriert und initialisiert. Der Zugriff auf die Datenbank erfolgt

vom Server selbst oder über einen Fernzugriff der erlaubten IP-Adressen. Der Datenbestand ist somit vor unbefugter Einsicht oder Manipulation gesichert.

Die Struktur der Forschungsdatenbank, also das Relationsschema, folgt größtenteils der Struktur der „ICUData“ Datenbank. Die Übernahme der Attributstruktur ermöglicht einen Fokus auf den wichtigen Schritt der Datenintegration in das System der Arbeitsgruppe. Aus technischer Sicht wäre für weitere Vermeidung an Datenreduanzen eine Umformung der Relationen in die 3. Normalform (siehe Kapitel 2.3.2) notwendig. Hierfür müssen die funktionalen Abhängigkeiten der Attribute analysiert werden. In diesem Bereich der Performance- und Speicheroptimierung besteht Verbesserungspotential, welches in zukünftigen Arbeiten weiter ausgeschöpft werden kann. Für die initiale Umsetzung des Projektes dieser Arbeit sind diese technischen Aspekte in Anbetracht der Arbeitsweise der Forschungsgruppe und der Menge an zu verarbeitenden Daten aber zunächst von zweitrangiger Priorität. In Bezug dazu muss auch berücksichtigt werden, dass eine weitere Aufteilung der Relationen zur Vermeidung von Datenredundanzen unweigerlich weitere Komplexität in die Struktur der Datenbank bringt, die alle weiteren Operationen wie Einfügen und Einsehen von Daten beeinflusst. Eine effektive Nutzung einer solchen Datenbank würde ein umfangreiches Wissen an Relationenalgebra und Datenbanktheorie voraussetzen, und somit den Zugang zu Informationen der klinischen Routinedaten erheblich erschweren. Es findet eine Abwägung zwischen einer ressourcensparenden, aber komplexeren Struktur und einem Aufbau statt, der die effektive Nutzung fördert. Da im Rahmen der Arbeit der Forschungsgruppe keine Echtzeit-Verarbeitung oder andere zeitkritische Vorgänge stattfinden, fällt die Entscheidung zugunsten der Implementierung einer simplifizierten Datenbankstruktur.

Die relationale Datenbank und dessen Struktur erfüllen die gestellte Anforderung **R3**. Die Anforderung **R2** wird durch die Nutzung der SQL-Sprache und dem in der Event Relation vorhandenen Attribut *Category* erfüllt. Ähnlich zu der in Bereich 9.1 erwähnten Flexibilität, kann die Datenbank mit neuen Attributen verändert und erweitert werden und kann somit auch an veränderte Anforderungen angepasst werden.

9.3 Zugriff und Nutzung

Der Zugriff auf die Daten kann durch die Nutzung des *psql* Kommandozeilen-Programms oder durch einen Fernzugriff mit Hilfe des *pgAdmin4* Programms erfolgen. Es ist allerdings auch möglich, jedes andere Datenbankmanagement-Programm, das mit PostgreSQL kompatibel ist, zu nutzen, da keine Beschränkungen von Seiten der Datenbank vorliegen. Für weitere Analysen oder automatisierte Verarbeitungen des Datenbestan-

des kann in einer Vielzahl an Programmier- und Skriptsprachen eine PostgreSQL Funktionsbibliothek importiert und eine Datenbankverbindung hergestellt werden. Somit steht die Datenbank für jede Softwarekomponente, die die Fähigkeit besitzt, eine Verbindung zu einer relationalen PostgreSQL Datenbank herzustellen, offen und es wird ein großer Freiraum an Möglichkeiten für die Nutzung der Daten in verschiedenen Szenarien geschaffen.

Die Auswahl und Ausgabe an Daten erfolgen über eine Kommunikation mit der Datenbank über SQL. Dies bietet ein einfaches aber mächtiges Werkzeug der Datenselektion und -verarbeitung mit der in Kapitel 2.5 beschrieben Funktionsweise und erfüllt so die Anforderung **R4**. Zu einer Veranschaulichung sind verschiedene Beispiele der möglichen Datenselektion mit SQL in den Abbildungen 9.1, 9.2 und 9.3 gegeben. Die in Abbildung 9.3 dargestellte SQL-Anfrage zeigt, welches Potential hinter einer Anfrage stecken kann. Dies ist allerdings keinesfalls notwendig, um effektiv Informationen aus der Datenbank zu extrahieren. So kann aber mit einer erworbenen Kompetenz in SQL die Verarbeitung von Daten aus der Datenbank um viele Schritte reduziert und vereinfacht werden.

```
SELECT *
FROM "Patient" INNER JOIN "Event" ON "Patient"."Pat_id" = "Event"."Pat_id"
LEFT OUTER JOIN "Value_nm" ON "Event"."Event_id" = "Value_nm"."Event_id"
LEFT OUTER JOIN "Value_tx" ON "Event"."Event_id" = "Value_tx"."Event_id"
WHERE "Observation_ts" BETWEEN '2000-01-01' AND '2000-02-02';
```

Abbildung 9.1: Selektion aller Attribute aller Patienten, denen Events in einem bestimmten Zeitraum zugewiesen sind.

```
SELECT DISTINCT "Patient"."Pat_id"
FROM "Patient" INNER JOIN "Event" ON "Patient"."Pat_id" = "Event"."Pat_id"
JOIN "Value_nm" ON "Event"."Event_id" = "Value_nm"."Event_id"
WHERE "Observation_id" like '%MDD_NBP-syst%' AND "Value_nm"."Value" > 100;
```

Abbildung 9.2: Selektion aller Patienten Pseudonyme, an denen eine systolische Blutdruckmessung durchgeführt wurde und der gemessene Wert größer als 100 ist.

```
SELECT *
FROM "Patient" INNER JOIN "Event" ON "Patient"."Pat_id" = "Event"."Pat_id"
LEFT OUTER JOIN "Value_nm" ON "Event"."Event_id" = "Value_nm"."Event_id"
LEFT OUTER JOIN "Value_tx" ON "Event"."Event_id" = "Value_tx"."Event_id"
WHERE "Patient"."Pat_id" IN
    (SELECT "Event"."Pat_id"
    FROM "Event" LEFT OUTER JOIN "Value_nm" ON
    "Event"."Event_id" = "Value_nm"."Event_id"
    WHERE "Observation_id" LIKE '%Resp_af%')
AND "Patient"."Pat_id" NOT IN
    (SELECT "Relationship".anchor_pseu
    FROM "Relationship"
    WHERE type LIKE "D")
AND "Patient"."Pat_id" NOT IN
    (SELECT "Relationship".pseu
    FROM "Relationship"
    WHERE type LIKE "D")
```

Abbildung 9.3: Verschachtelte Selektion aller Attribute aller Patienten Pseudonyme, von Patienten, denen eine Atemfrequenz Event zugeordnet ist, ohne Patienten, die sich Patienten IDs mit anderen Personen teilen.

9.4 Schlussfolgerung

Die in Kapitel 6.3 gestellten Anforderungen **R1** bis **R5** wurden mit der Konzeption und Implementierung eines funktionsfähigen Systems der Datenintegration und -speicherung von klinischen Routinedaten aus der KIS Datenbank der Klinik für Neurologie des Universitätsklinikums Magdeburg erfüllt. Darüber hinaus wurde mit der Inklusion vieler Optionen und Konfigurationsparameter in jedem Schritt eine Flexibilität in der Datenverarbeitung erreicht, die eine langfristige Nutzung der erstellten Skripte und Datenbank ermöglicht.

10 Fazit und Ausblick

In dieser Arbeit wurden die verschiedenen Aspekte der digitalen Datenverwaltung der medizinischen Domäne beleuchtet. Es wurde eine Datenbank für die Arbeitsgruppe MedDigit für klinische Routinedaten der Klinik für Neurologie der Universitätsklinik Magdeburg entwickelt und implementiert, die das Spektrum an für Forschungszwecke verfügbaren medizinischen Daten erweitern soll. Der Prozess der Datenbankerstellung und Datenintegration wurden erfolgreich durchgeführt und erfüllen die grundlegenden Anforderungen, die an ein solches System gestellt wurden.

10.1 Fazit

Der erstellte Prozess der Datenintegration geht einen ersten Schritt der Exploration von bisher unverfügbaren Daten der klinischen Patientenversorgung. Die Analyse dieser Daten gibt die Möglichkeit der Generierung und Bestätigung neuer Hypothesen im Rahmen der medizinischen Forschung. Die Integration klinischer Routinedaten in ein für die Forschung nutzbares IT-System ist ein klarer Fortschritt in der weiteren Digitalisierung und Vernetzung von Forschung, Wissenschaft und dem Gesundheitswesen.

Die Entwicklung eines Datenspeichers für medizinische Daten stellt in dem Feld der Informatik und der Medizin kein Novum dar, jedoch war die Implementierung mit einer Reihe an neuen, sich aus den lokalen Gegebenheiten ergebenden Herausforderungen verbunden. Es musste zunächst eine Interpretation des vorhandenen Datenbestandes stattfinden, da das KIS einer eigenen proprietären Implementierung der Datenverwaltung folgt. So hat eine nicht eindeutige Zuweisung von Patientenidentifikatoren dieser Implementierung grundlegende Probleme in der Umsetzung dieser Arbeit verursacht. Die Analyse dieser Besonderheiten der Krankenhausdatenbank und die bisher gefundenen Lösungsansätze sind neue Erkenntnisse, die für zukünftige Arbeiten und Entwicklungen in diesem Zusammenhang wertvoll sind.

Das Ergebnis dieser Arbeit bietet noch Verbesserungs- und Weiterentwicklungspotential, aber der erstellte Datenintegrationsprozess und die Datenbank gehen über eine rein theoretische oder prototypische Entwicklung hinaus. Das erstellte System kann in seiner derzeitigen Variante Daten aus dem Export der KIS Datenbank der Klinik für Neurologie einlesen, filtern, anonymisieren, unter den gegeben Funktionen des Mainzelliste-Programms pseudonymisieren und diese verarbeiteten Daten in eine für diesen Zweck erstellte Datenbank einfügen. Dies macht eine Inkorporation von klinischen Routinedaten in die Forschungsarbeiten der Arbeitsgruppe MedDigit möglich.

10.2 Ausblick

Die Datenbank bietet eine Reihe an potenziellen Weiterentwicklungen sowohl in dem Prozess der Datenintegration, der Anonymisierung und Pseudonymisierung und der Datenbank selbst. Zunächst kann eine weitere Analyse des Datenbestandes der Klinik für Neurologie durchgeführt werden, um weitere Informationen über die Struktur und Inhalt der noch nicht eingebundenen Attribute der obx und obr Dateien zu erhalten. Eine Einbindung weiterer Attribute und Daten ist aufgrund der bereits implementierten Konfigurationsparameter des Datenverarbeitungsprogramms mit geringem Aufwand möglich und ermöglicht so auf einem unkomplizierten Weg, Daten aus der KIS Datenbank einzubinden.

Eine zusätzliche Möglichkeit, der Datenbank neue Informationen zuzuführen, ist die Einbindung von Freitext-Daten in den Datenbestand. Da im Rahmen dieser Bachelorarbeit keine Prozessierung dieser Daten möglich ist, werden diese, wie in den vorherigen Kapiteln beschrieben, in der Datenintegration ausgefiltert, um keine persönlichen Patienteninformationen in die Forschungsdatenbank aufzunehmen. Durch ein *Natural Language Processing* könnte der Inhalt von Freitext-Daten analysiert und kategorisiert werden. Somit können gezielt identifizierende Informationen herausgefiltert und neue Daten und Events in die Datenbank eingefügt werden, die viele Freitext-Eintragungen besitzen, wie z.B. ärztliche Diagnosen und Untersuchungen.

Der Datenbestand kann mit einer Standardisierung und Harmonisierung mit anderen medizinischen Datenbeständen kompatibel und vergleichbar gemacht werden. Eine Integration in neue medizinische Datenstandards kann eine Interoperatibilität der Daten der Magdeburger Klinik für Neurologie mit anderen nationalen und internationalen Datenbeständen herstellen.

A

Abbildungsverzeichnis

B

Quellenverzeichnis

[1] Kassenärztliche Bundesvereinigung - Bundesärztekammer, “Hinweise und Empfehlungen zur ärztlichen Schweigepflicht, Datenschutz und Datenverarbeitung in der Arztpraxis,” *Deutsches Ärzteblatt*, vol. 115, no. 10, p. 453, März 2018.

[2] Branchenführer 2014 Healthcare IT, *Interview mit Prof. Dr. Peter Haas*. E Health Compendium, 2014.

[3] M. Dugas, K.-H. Jöcke, O. Gefeller, P. Knaup-Gregori, T. Friede, E. Ammenwerth, M. Kieser, and H.-U. Prokosch, “Freier Zugang zu Dokumentationsformularen und Merkmalskatalogen im Gesundheitswesen. Memorandum „Open Metadata“,” *GMS Medizinische Informatik, Biometrie und Epidemiologie*, vol. 10, 2014.

[4] O. Rienhoff and S.C. Semler, *Schriftenreihe der TMF, Terminologien und Ordnungssysteme in der Medizin*. Medizinisch Wissenschaftliche Verlagsgesellschaft, September 2015, no. 13.

[5] Medizinische Fakultät Universitätsklinikum Magdeburg, Forschungsgruppe Medizin und Digitalisierung, “Medizin und Digitalisierung.” [Online]. Available: http://www.kneu.ovgu.de/MedDigit.html

[6] Oracle Coperation, “Datenbank Definition.” [Online]. Available: https://www.oracle.com/database/what-is-database.html

[7] M. Rouse, “Datenbank Definition.” [Online]. Available: https://searchsqlserver.techtarget.com/definition/database

[8] G. Saake, K.-U. Sattler, and A. Heuer, *Datenbanken - Konzepte und Sprachen*, 4th ed. MITP-Verlag, 2010.

[9] A. Autor, “Die neun Regeln von Codd und ein Einblick in das formale Relationenmodell.” [Online]. Available: https://blog.triona.de/development/database/die-neun-regeln-von-codd-und-ein-einblick-in-das-formale-relationenmodell.html

[10] E. Codd, “Relational database: a practical fundation for productivity,” *Communications of the ACM*, vol. 25, no. 2, pp. 109–117, Februar 1982.

[11] C. Date, “A.M. Turing Award - Edgar F. Codd.” [Online]. Available: https://amturing.acm.org/award_winners/codd_1000892.cfm

[12] “Excel specifications and limits.” [Online]. Available: https://support.office.com/en-us/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3?ui=en-US&rs=en-US&ad=US

[13] M. L. Brodie and J. W. Schmidt, “Final Report of the ANSI/X3/SPARC DBS-SG Relational Database Task Group, Doc. No. SPARC-81-690,” September 1981.

[14] “Three Level Database Architecture.” [Online]. Available: http://www.dbmsinternals.com/database-fundamentals/basic-architecture/three-level-database-architecture/

[15] B. Lambeau, “About Logical Data Independence (2).” [Online]. Available: http://www.revision-zero.org/logical-data-independence-2

[16] “ANSI-3-Ebenenmodell Definition.” [Online]. Available: https://www.datenbanken-verstehen.de/lexikon/ansi-drei-ebenenmodell/

[17] C. Tupper, *Data Architecture - From Zen to Reality*. Elsevier, 2011.

[18] T. Studer, *Relationale Datenbanken*, 2nd ed. Springer-Verlag, 2019.

[19] S. Bagui and R. Earp, *Database Design Using Entity-Relationship Diagrams*, 2nd ed. Boca Raton, Fla: CRC Press, 2011.

[20] “Entity-Relationship-Modell Definition.” [Online]. Available: https://www.datenbanken-verstehen.de/lexikon/er-diagramm/

[21] I. Bouchrika, “How to Convert ER Diagram to Relational Database.” [Online]. Available: https://www.learndb.com/databases/how-to-convert-er-diagram-to-relational-database

[22] E. Schicker, *Datenbanken und SQL - Eine praxisorientierte Einführung mit Anwendungen in Oracle, SQL Server und MySQL*, 5th ed. Springer-Verlag, 2017.

[23] International Organization for Standardization, “ISO 9075:1987 Information processing systems — Database language — SQL.” [Online]. Available: https://www.iso.org/standard/16661.html

[24] aerzteblatt.de, “SNOMED CT-Pilotlizenz: Meilenstein für die Standardisierung.” [Online]. Available: https://www.aerzteblatt.de/nachrichten/111154/SNOMED-CT-Pilotlizenz-Meilenstein-fuer-die-Standardisierung

[25] H. E. Krüger-Brand, “Meilenstein für die standardisierung,” *Deutsches Ärzteblatt*, vol. 117, no. 15, pp. 654–655, April 2020.

[26] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), “Klassifikation ICD-10-GM.” [Online]. Available: https://www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-gm/

[27] American Health Information Management Association (AHIMA), “ICD-10 Primer.” [Online]. Available: https://library.ahima.org/doc?oid=106177

[28] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), “Aufbau der [...] Semantik der ICD-10-GM.” [Online]. Available: https://www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-gm/systematik/systematik/

[29] Deutsches Institut für Medizinische Dokumentation und Information im Auftrag des Bundesministeriums für Gesundheit, “Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme - German Modification (ICD-10-GM) Systematisches Verzeichnis,” 2019.

[30] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), “Klassifikation OPS.” [Online]. Available: https://www.dimdi.de/dynamic/de/klassifikationen/ops/

[31] RI Innovation GmbH, Reimburstment Institute, “Glossar eintrag OPS.” [Online]. Available: https://reimbursement.institute/glossar/prozedur/

[32] Deutsches Institut für Medizinische Dokumentation und Information im Auftrag des Bundesministeriums für Gesundheit, “Operationen- und Prozedurenschlüssel Internationale Klassifiaktion der Prozeduren in der Medizin (OPS) Systematisches Verzeichnis,” 2019.

[33] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), "Klassifikation ATC/DDD." [Online]. Available: https://www.dimdi.de/dynamic/de/arzneimittel/atc-klassifikation/

[34] Deutsches Institut für Medizinische Dokumentation und Information im Auftrag des Bundesministeriums für Gesundheit, "Amtliche Fassung des ATC-Index mit DDD-Angaben für Deutschland im Jahre 2020," 2020.

[35] SNOMED International, *SNOMED CT Starter Guide*. International Health Terminology Standards Development Organisation, 2017.

[36] American Medical Association, "CPT overview and code approval." [Online]. Available: https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval

[37] American Academy of Professional Coders, "CPT Codes." [Online]. Available: https://coder.aapc.com/cpt-codes

[38] American Medical Association, "Healthcare Common Procedure Coding System (HCPCS)." [Online]. Available: https://www.ama-assn.org/practice-management/cpt/healthcare-common-procedure-coding-system-hcpcs

[39] LabCE by MediaLab, "HCPCS and CPT-4 Coding." [Online]. Available: https://www.labce.com/spg257613_hcpcs_and_cpt_coding.aspx

[40] American Academy of Professional Coders, "Structure of Level II HCPCS Codes." [Online]. Available: https://www.aapc.com/resources/medical-coding/hcpcs.aspx#StructureofLevelIIHCPCSCodes

[41] ——, "HCPCS Codes." [Online]. Available: https://coder.aapc.com/hcpcs-codes

[42] Regenstrief Institute, "Origins of LOINC." [Online]. Available: https://loinc.org/about/

[43] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), "Klassifikation LOINC und RELMA." [Online]. Available: https://www.dimdi.de/dynamic/de/klassifikationen/weitere-klassifikationen-und-standards/loinc-relma/

[44] Regenstrief Institute, "LOINC CODE 2339-0." [Online]. Available: https://loinc.org/2339-0/

[45] National Library of Medicine (NIH), “RxNorm Overview.” [Online]. Available: https://www.nlm.nih.gov/research/umls/rxnorm/overview.html

[46] National Center for Biomedical Ontology, “RxNorm Overview.” [Online]. Available: http://purl.bioontology.org/ontology/RXNORM/18631

[47] Observational Health Data Sciences and Informatics, “Data Standardization OMOP Common Data Model.” [Online]. Available: https://www.ohdsi.org/data-standardization/

[48] ——, “The Book of OHDSI,” April 2020.

[49] Cincinnati Children’s Hospital Medical Center, “Informatics for Integrating Biology and the Bedside (i2b2).” [Online]. Available: https://i2b2.cchmc.org/faq

[50] Deutscher Ethikrat, *Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung*, November 2017.

[51] Prof. Dr. Dougas, Martin, “Smart Medicine: Im Zeitalter von Big Data mehr als eine Vision.” [Online]. Available: https://blog.der-digitale-patient.de/big-data-debattenreihe-dugas/

[52] D. B. Taichman, J. Backus, C. Baethge, H. Bauchner, P. W. de Leeuw, J. M. Drazen, J. Fletcher, F. A. Frizelle, T. Groves, A. Haileamlak, A. James, C. Laine, L. Peiperl, A. Pinborg, P. Sahni, and S. Wu, “Sharing clinical trial data: a proposal from the international committee of medical journal editors,” *The Lancet*, vol. 387, no. 10016, Januar 2016.

[53] ——, “Sharing clinical trial data: a proposal from the international committee of medical journal editors,” *The New England Journal of Medicine*, vol. 374, no. 4, Januar 2016.

[54] Deutsches Institut für Medizinische Dokumentation und Information (DIMDI), “G-DRG-System - Fallpauschalen in der stationären Versorgung.” [Online]. Available: https://www.dimdi.de/dynamic/de/klassifikationen/ops/anwendung/zweck/

[55] RI Innovation GmbH, Reimburstment Institute, “Glossar Eintrag G-DRG System.” [Online]. Available: https://reimbursement.institute/glossar/prozedur/

[56] Bundesministerium für Bildung und Forschung (BMBF), “Medizininformatik.” [Online]. Available: https://www.gesundheitsforschung-bmbf.de/de/medizininformatik.php

[57] Medizininformatik-Initiative (MI-I), “Über die Initiative.” [Online]. Available: https://www.medizininformatik-initiative.de/de/ueber-die-initiative

[58] ——, “FAQ.” [Online]. Available: https://www.medizininformatik-initiative.de/de/ueber-die-initiative/faq-haeufig-gestellte-fragen

[59] H. E. Krüger-Brand, “Medizininformatik-Initiative: Impulse für die digitale Medizin,” *Deutsches Ärzteblatt*, vol. 116, no. 42, p. 1883, Oktober 2019.

[60] Universitätsmedizin Mainz, “Die Mainzelliste - Pseudonymisierung und Identitätsmanagement.” [Online]. Available: https://www.unimedizin-mainz.de/imbei/informatik/ag-verbundforschung/mainzelliste.html

[61] M. Lablans, A. Borg, and F. Ückert, “A RESTful interface to pseudonymization services in modern web applications,” *BMC Medical Informatics and Decision Making*, vol. 15, no. 2, Februar 2015.

[62] A. Faldum and K. Pommerening, “An optimal code for patient identifiers.” *Computer methods and programs in biomedicine*, vol. 79, no. 1, pp. 81–8, Jul 2005.

[63] H.-U. Prokosch *et al.*, “MIRACUM: Medical Informatics in Research and Care in University Medicine.” *Methods of Information in Medicine*, vol. 57, no. S 01, pp. e82–e91, Juli 2018.

[64] Medical Informatics in Research and Care in University Medicine, “MIRACOLIX Tools.” [Online]. Available: https://www.miracum.org/miracolix-tools/

[65] T. Ganslandt, C. Haverkamp, H. Storf, and H.-U. Prokosch, “MIRACUM Live: Forschungs- und Patientenportal,” Youtube. [Online]. Available: https://youtu.be/mmQpN_Jtg6g

[66] H.-U. Prokosch, “Kernkomponenten und ETL-Prozesse in einem MIRACUM DIZ,” Youtube. [Online]. Available: https://youtu.be/gISD_y_9Xx0

[67] Institut für das Entgeltsystem im Krankenhaus InEK GmbH, “Datenlieferung gem. § 21 KHEntgG.” [Online]. Available: https://www.g-drg.de/Datenlieferung_gem._21_KHEntgG

[68] Medical Informatics in Research and Care in University Medicine, “MIRACUM Publications.” [Online]. Available: https://www.miracum.org/further-readings/publications/

[69] FileInfo, ".TXT File Extension." [Online]. Available: https://fileinfo.com/extension/txt

[70] D. Dr. med. Brammen, S. Dr.-Ing. Oeltze-Jafra, J. Müller, and J. Schmidt, "Protokoll zum Treffen mit Dr. Dominik Brammen vom 10.03.2020."

[71] Rote Liste Service GmbH, "Über die ROTE LISTE." [Online]. Available: https://www.rote-liste.de/ueber-rote-liste

[72] The PostgreSQL Global Development Group, "PostgreSQL." [Online]. Available: https://www.postgresql.org/about/

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Bachelorarbeit selbstständig, nur unter Zuhilfenahme der aufgeführten Quellen und Hilfsmittel, verfasst zu haben. Diese Arbeit wurde weder einer anderen Prüfungsbehörde vorgelegt noch veröffentlicht.

Datum: .....

(Unterschrift)