Projekt „QI-KA“

„Sektorübergreifende kardiologische Qualitätsindikatoren für das österreichische Gesundheitswesen“

Bericht Datenaufbereitung

Arbeitsgemeinschaft Medizinische Universität Wien

CeMSIIS, Institut für Klinische Biometrie

Georg Heinze, Ao. Univ.-Prof. Dr.
Hana Sinkovec, Mag. Univ.-Dipl.

Universitätsklinik für Innere Medizin 2, Klinische Abteilung für Kardiologie

Alexander Niessner, Univ.-Prof. Dr. med.
Patrick Sulzgruber, Dr. med.

CeMSIIS, Institut für Medizinisches Informationsmanagement

Walter Gall, Ao. Univ.-Prof. Dr.
Florian Katsch, BSc.
Marko Todorovic, BSc.
Georg Duftschmid, Univ.- Prof. Dr.

Projektleitung

Ao. Univ.-Prof. Dr. Walter GALL
walter.gall@meduniwien.ac.at

Institut für Medizinisches Informationsmanagement
Zentrum für Medizinische Statistik, Informatik und Intelligente Systeme
Medizinische Universität Wien, Spitalgasse 23, A-1090 Wien

Wien, 22.05.2019

Inhaltsverzeichnis

1 Einleitung

Anmerkung zur geschlechtergerechten Formulierung: im Interesse einer besseren Lesbarkeit wird nicht ausdrücklich in geschlechtsspezifischen Personenbezeichnungen differenziert. Die gewählte männliche Form schließt eine adäquate weibliche Form gleichberechtigt ein.

1.1. Projekt QI-KA

Im Fokus des Projektes steht die Entwicklung von Indikatoren für die Qualität des österreichischen Gesundheitswesens, wobei stationäre und niedergelassene Behandlungsprozesse gemeinsam betrachtet werden. Die Ergebnisse sollen unter anderem Aufschluss über die leitliniengerechte Behandlung, insbesondere von Herzinfarkten, geben und als Diskussionsgrundlage in regionalen Versorgungszonen dienen.

1.2. Dokumentbeschreibung

Das vorliegende Dokument beschreibt die schrittweise Anpassung der importierten Rohdaten in ein zuverlässiges, ausreichend performantes und annotiertes Datenmodell sowie die semantische Beschreibung der darin enthaltenen Daten und das Hinzufügen von zusätzlichem Datenmaterial. Die beschriebenen Arbeitsschritte sind entsprechend der Projektbeschreibung wiederverwendbar, in den benannten SQL Sourcen einsehbar und werden in diesem Dokument ausgewiesen.

Weiters werden vorab getätigte statistische Analysen angeführt um die verfügbaren Daten rudimentär zu beschreiben. Hierbei liegt der Fokus auf einer quantitativen Beschreibung der vorhandenen Daten. Zu jeder Tabelle im Datenschema folgt nach einer Beschreibung, eine Erklärung der getätigten Anpassung, eine Übersichtsstatistik und die Abfrage von 5 zufällig ausgewählten Einträgen um die Daten zu veranschaulichen. Hierbei sind die Spalten, aus welchen der Primary Key besteht jeweils kursiv angeschrieben. Danach folgen je nach Tabelle unterschiedliche Auswertungen auf den Daten. Eine kurze Beschreibung der durchgeführten Abfrage ist direkt vor dem jeweiligen verwendeten SQL-Code verfügbar. Hinweis: Manche der Tabellen mussten aufgrund ihrer Breite auf mehrere Zeilen aufgespalten werden, sie sind mit “Table continues below” markiert.

2 Importierte Rohdaten

Die Rohdaten stammen vom Hauptverband der Österreichischen Sozialversicherungsträger und werden aus der Forschungsdatenbank LEICON bezogen. Alle Relationen stellen Abrechnungsdaten der Sozialversicherungen dar. Diese Daten beinhalten Stammdaten sowie Leistungen aus dem niedergelassenen Bereich und Apotheken. Die drei Relationen mit dem Präfix mbds stammen aus den Meldungen an das Bundesministerium für Gesundheit und Frauen und beinhalten Daten zu Spitalsaufenthalten inklusive Diagnosen und Einzelleistungen entsprechend der LKF. Das Datenset enthält Abrechnungsdaten von, seitens der Auftraggeber vorselektierten, Patienten mit Herzinfarkt aus dem niedergelassenen und stationären Bereich aus den Jahren 2011 bis 2016.

In Abbildung 1 werden die einzelnen Relationen der Rohdaten dargestellt. Im Folgenden werden diese Relationen genauer beschrieben.

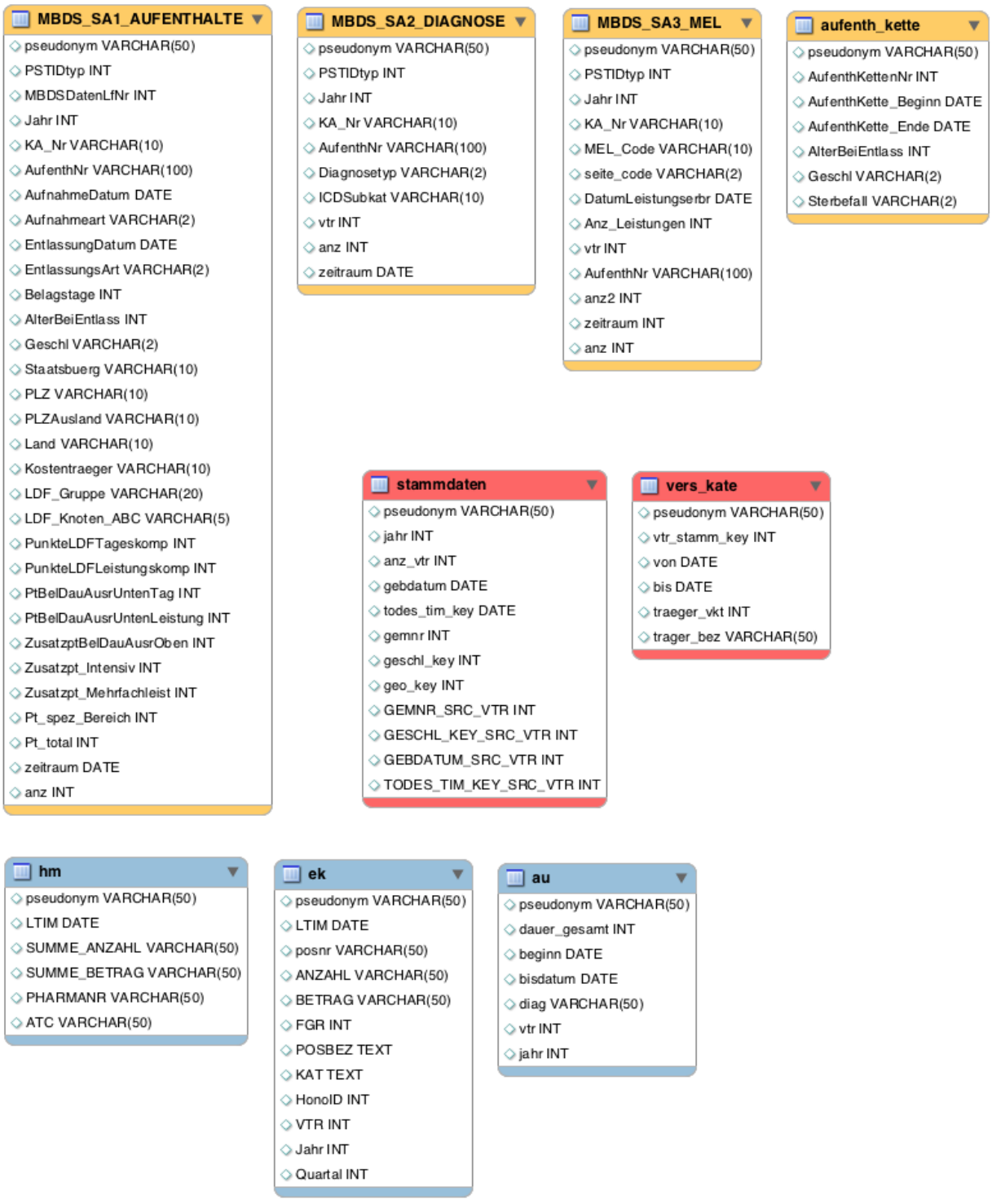

Abbildung 1: Überblick über die Relationen der Rohdaten

Darin enthalten sind Relationen aus der MBDS Datenquelle (orange), Relation zur Beschreibung der Patienten aus dem Leicon Datensatz stammend (rot). Ebenfalls auf dem Leicon Datensatz stammen Relationen zu Heilmittel, Leistungen und Arbeitsunfähigkeit (blau).

2.1 Tabelle mbds_sa1_aufenthalte

Die Daten dieser Tabelle stammen aus den Meldungen der Krankenanstalten, dem sog. MBDS (Minimum Basic Data Set), an das Bundesministerium für Gesundheit und Frauen

(Sozialministerium) im Rahmen der leistungsorientierten Krankenanstaltenfinanzierung (LKF). Es handelt sich um die Leistungs- und Diagnosedaten der österreichischen Fondskrankenanstalten und teilweise weiterer Krankenanstalten. Die MBDS Daten werden vom Sozialministerium in anonymisierter Form (keinerlei Personenbezug) zur Verfügung gestellt und wurden vom Auftraggeber mittels eines nicht näher bekannten Matching-Verfahrens den weiteren Daten zugeordnet.

Die MBDS Datensätze werden zur Berechnung der LKF-Fallpauschale seitens des Sozialministeriums herangezogen und beinhalten Datensätze der landesgesundheitsfondsfinanzierten öffentlichen Krankenanstalten und Ordensspitäler, sowie teilweise Privatkrankenanstalten-Finanzierungsfonds finanzierte Krankenanstalten.

Die MBDS Daten sind im vorliegenden Schema in drei Relationen geteilt, wobei die Erste behandelte Krankenhausaufenthalte abbildet und die beiden anderen dazu zugeordnete Diagnosen bzw. medizinische Einzelleistungen (MEL).

Die Relation mbds_sa1_aufenthalte enthält 241.446 Datensätze von 47.166 verschiedenen Patienten über die Jahre 2011 bis 2016.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	identifiziert einen Patienten über alle Relationen
pstidtyp	int - enthält nur die Werte null und 3	unbekannte Bedeutung. Zur Diskussion steht die Bedeutung als Maß der Anonymität der einzelnen Datensätze. Dieses Attribut scheint jedoch nicht von Bedeutung für die gegenständlichen Analysen
mbdsdatenlfnr	int - eindeutige Nummer für einen Aufenthalt im jeweiligen Berichtsjahr	-
jahr	int - Berichtszeitraum	-
ka_nr	varchar(10) - Eindeutige Identifikationsnummer einer Krankenanstalt	vgl. GAP-DRG Wiki [1]
aufenthnr	varchar(100) - Eindeutige Identifikation des Aufenthaltes über alle Datensätze innerhalb eines Krankenhauses	lt. explorativer Analyse sind die Werte in `aufenthnr` eindeutig per Krankenhaus, im MBDS Datensatz als `AUFENTHALT_NR` codiert, vgl. GAP-DRG Wiki [1]
aufnahmedatum	date - Datum der Aufnahme	Das Aufnahmedatum kann neben dem Datum der Aufnahme des/der Patienten/in das Krankenhaus

		auch das Datum der krankenhausinternen Verlegung in einen anderen Leistungsbereich (z.B. in den ausschließlichen Bereich der Pflege) oder das Datum des Eintritts der Asylierung (Datum, ab dem der stationäre Krankenhausfall von der Sozialversicherung nicht mehr als Krankheitsfall anerkannt wurde) sein (vgl. hierzu GAP-DRG Wiki [1]).
aufnahmeart	varchar(2) - Bezeichnet die Art der Aufnahme und gleichzeitig den Leistungsbereich, in welchem der stationäre Patient aufgenommen wurde	(vgl. GAP-DRG Wiki [1]) - A - Aufnahme in den allgemein stationären Bereich inkl. Aufnahme auf Intensiveinheiten - T - Transfer von einem anderen Krankenhaus in den allgemein stationären Bereich inkl. Intensiveinheiten - W - Wiederaufnahme in den allgemein stationären Bereich inkl. Wiederaufnahme auf Intensiveinheiten - R - Aufnahme in den Bereich der Rehabilitation - P - Aufnahme in den ausschließlichen Bereich der Pflege - H - Aufnahme in den halbstationären Bereich - L - Kennzeichen der Datensätze, die den Patientenaufenthalt nach dem Zeitpunkt der Asylierung beschreiben- K - Aufnahme eines 0-Tagesfalls
entlassungdatum	date - Datum der Entlassung	Dieses Datum kann neben dem Zeitpunkt der Entlassung, des Todes oder des Transfers des Patienten ein anderes Krankenhaus auch der Zeitpunkt der krankenhausinternen Verlegung in einen anderen Leistungsbereich (z.B. in den ausschließlichen Bereich der Pflege) oder der Zeitpunkt des Eintritts der Asylierung (Zeitpunkt, zu dem der

		stationäre Krankenhausfall von der Sozialversicherung nicht mehr als Krankheitsfall anerkannt wurde) sein. (vgl. Attribut `aufnahmedatum`, vgl. GAP-DRG Wiki [1])
entlassungsart	varchar(2) - Art der Entlassung	(vgl. GAP-DRG Wiki [1])- E Entlassung aus dem Krankenhaus- T Transfer in ein anderes Krankenhaus - S Sterbefall - A Krankenhausinterne Verlegung vom Bereich der Rehabilitation und vom ausschließlichen Bereich der Pflege in den allgemeinen stationären Bereich (inkl. Intensivbereich) - H Abschluss eines Aufenthaltes im halbstationären Bereich oder krankenhausinterne Verlegung in den halbstationären Bereich - R Krankenhausinterne Verlegung in den Bereich der Rehabilitation - P Krankenhausinterne Verlegung in den ausschließlichen Bereich der Pflege - L Kennzeichen der Datensätze, die zum Zeitpunkt der Asylierung dokumentarisch abgeschlossen werden - V Kennzeichen für noch nicht abgeschlossene Aufenthalte von am Jahresende verbleibenden Patienten - 4 Entlassung gegen Revers (lt. Rückfrage mit NÖGKK)
belagstage	int - Summer der Mitternachtsstände	Belagstage des Aufenthaltes. Dies ist eine redundante Information da durch die Attribute `aufnahmedatum` und `entlassungsdatum` berechenbar.
alterbeientlass	int - Alter des Patienten bei Entlassung in Jahren	(vgl. Relation `stammdaten`)

geschl	varchar(2) - Geschlecht des Patienten	codiert als 'm'/'w' (vgl. Relation `stammdaten`)
staatsbuerg	varchar(10) Ländercode lt. Sozialministerium, bezeichnet die Staatsbürgerschaft eines Patienten	(vgl. Relation `stammdaten`, teilweise alternierende `staatsbuerg` für Patienten gefunden)
plz	varchar(10) - Hauptwohnsitz des Patienten (österreichische PLZ lt. Systematik des Sozialministeriums oder Wert "AUSL")	-
plzausland	varchar(10) - siehe Hauptwohnsitz, bei Hauptwohnsitz im Ausland (optional)	-
land	varchar(10) - Wohnsitzland im Ausland (optional), Ländercodes lt. Sozialministerium	-
kostentraeger	varchar(10) - Kostenträgercode (numerisch), pro Aufenthalt ist exakt ein Kostenträger bekannt;	Datensatz enthält neben numerischen Werten auch alphanummerische Werte welche gesonderte Kostenträger notieren (z.B. Länder), die Kostenträger werden in weiterer Folge in eigener Relation beschrieben und benannt
ldf_gruppe	varchar(20) - Fallpauschalengruppe (lt. LKF) (= Leistungsorientierte Diagnosenfallgruppe), z.B. 'MEL01.01'	Bei Datensätzen von stationären Krankenhausaufenthalten, die keiner leistungsorientierten Diagnosenfallgruppe zugeordnet werden, erfolgt ein Eintrag entsprechend der jeweiligen Aufnahmeart: - DIAG0PKT - bei Patienten, deren codierte Hauptdiagnose für einen stationären Aufenthalt alleine nicht plausibel ist - FEHLER - bei fehlerhaften Datensätzen - GERIAT - bei Patienten mit ausschließlichem Aufenthalt

		auf einer Einheit für Akutgeriatrie/Remobilisation - KJP- bei Patienten/-innen mit ausschließlichem Aufenthalt auf einer Einheit für Kinder- und Jugendpsychiatrie - LANGZEIT - bei der Aufnahmeart „L“ - NEURO - bei Patienten mit ausschließlichem Aufenthalt auf einer Einheit für Akut-Nachbehandlung von neurologischen Patienten - PALLIAT - bei Patienten mit ausschließlichem Aufenthalt auf einer palliativmedizinischen Einheit - PFLEGE - bei der Aufnahmeart „P“ - REMOB - bei Patienten mit ausschließlichem Aufenthalt auf einer Einheit für Remobilisation/Nachsorge - VERBLEIB - bei am Jahresende Verbleibenden Patienten
ldf_knoten_abc	varchar(5) - Fallpauschale (A-F) falls reguläre Fallgruppe vorhanden	Weitere Werte (‘-’, ‘G’-‘L’) sind enthalten, wenn beispielsweise spezielle Fallpauschalengruppe angegeben wurden.
punkteldftageskomp	int - Punkte für die Tageskomponente lt. LDF (Hotelkomponente)	-
punkteldfleistungskomp	int - Punkte für die Leistungskomponente lt. LDF (erbrachte medizinische Einzelleistungen)	-
ptbeldauausruntentag	int - Ausreißer Referenz-Belagsdauer unterschritten (Pkt.)	-
ptbeldauausruntenleistung	int - Ausreißer Referenz-Leistung unterschritten (Pkt.)	-
zusatzptbeldauausroben	int - Ausreißer Referenz-Belagsdauer überschritten (Pkt.)	-

zusatzpt_intensiv	int - Zusatzpunkte Intensiveinheiten (Pkt.)	-
zusatzpt_mehrfachleist	int - Zusatzpunkte Mehrfachleistungen (Pkt.)	-
pt_spez_bereich	int - Zusatzpunkte spezielle Bereiche (Pkt.)	z.B. in den Bereichen der Kinder- und Jugendpsychiatrie, der Akut-Nachbehandlung von neurologischen Patienten, der medizinischen Geriatrie, der Akutgeriatrie/Remobilisation sowie der palliativmedizinischen Einrichtungen
pt_total	int - Gesamtpunkte des Aufenthaltes (Pkt.)	-
zeitraum	date (40 distinkte Werte, > 50% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben
anz	int (1 distinkter Wert '1', >55% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben

2.2 Tabelle stammdaten

Die Relation stammdaten beschreibt Attribute zugehörig zu einem Patienten. Die Datensätze sind pro Jahr und pro Patient eindeutig, es existieren also für jedes Jahr Daten zu einem Patienten. Es sind neben den grundlegenden Daten zu einer Person auch die jeweiligen Ursprünge der Daten festgehalten; die Attribute mit dem Postfix '_src_vtr' im Attributnamen enthalten diejenigen Versicherungsträger welche diesen Datensatz erhoben haben und beschreiben so den Ursprung der jeweiligen Daten.

Es sind 255.505 Datensätze von 46.235 Patienten von den Jahren 2011 bis 2016 enthalten.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	Mehrere Datensätze pro Patient.
jahr	int - Jahr des Datensatzes	-
anz_vtr	int - Anzahl der Versicherungsträger	Als Bedeutung wird die Anzahl der Versicherungsträger in diesem Jahr die einem Pseudonym zugeordnet wurden angenommen. Da der vorliegende Datensatz nicht vollständig ist kann diese

		Annahme nicht überprüft werden.
gebdatum	date - Geburtsdatum	-
todes_tim_key	date - Todesdatum	-
gemnr	int - Gemeindecode (GCD) oder Gemeinde Kennziffer lt. Adressregisterverordnung	Siehe Statistik Austria [2]
geschl_key	int - Geschlecht	Geschlecht wird als Zahl kodiert (1=männlich, 2=weiblich), eine erste Validitätsprüfung zeigt Pseudonyme mit mehrfacher Geschlechterzuordnung.
geo_key	int - Gemeindecode (GCD) oder Gemeinde Kennziffer lt. Adressregisterverordnung	siehe Statistik Austria [2]. Die Gemeindecodes bzw. Gemeindekennziffer sind österreichweit im Wesentlichen ident. Lediglich in Wien (Gemeindekennziffer 90001 und Unterteilungen entsprechend den Bezirken) unterscheiden sich die beiden Systeme. Im vorliegenden Attribut sind beide Codiersysteme gemischt vorhanden. Eine Bereinigung dieser Daten ist jedoch trivial.
gemnr_src_vtr	int - Versicherungsträger (siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)	Herkunft der Daten im Attribut gemnr
geschl_key_src_vtr	int - Versicherungsträger (siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)	Herkunft der Daten im Attribut geschl_key
gebdatum_src_vtr	int - Versicherungsträger (siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)	Herkunft der Daten im Attribut gebdatum
todes_tim_key_src_vtr	int - Versicherungsträger (siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)	Herkunft der Daten im Attribut todes_tim_key

2.3 Tabelle mbds_sa2_diagnose

Diese Tabelle ist in enger Beziehung zur zuvor Beschriebenen zu sehen und enthält die Haupt und Nebendiagnosen zugehörig zu einem Krankenhausaufenthalt. Die Relation

mbds_sa2_diagnose enthält 870.869 Datensätze von 47.166 verschiedenen Patienten (stimmt mit Anzahl an Aufenthalten überein) über die Jahre 2011 - 2016.

Attribut	Inhalt	Bemerkung
pseudonym	siehe `mbds_sa1_aufenthalte`	-
pstidtyp	siehe `mbds_sa1_aufenthalte`	-
jahr	siehe `mbds_sa1_aufenthalte`	-
ka_nr	siehe `mbds_sa1_aufenthalte`	-
aufenthnr	siehe `mbds_sa1_aufenthalte`	-
diagnosetyp	varchar(2) - Haupt- oder Zusatzdiagnose (H, Z)	-
icdsubkat	varchar(10) - Diagnose codiert	Die Daten enthalten ICD-10 codierte Diagnosen verschiedener Versionen.
vtr	int - Versicherungsträger	Versicherungsträger dem diese Diagnose zugeordnet ist als nummerischer Code. Die Daten dieses Attributs enthalten ausschließlich Versicherungsträger (im Gegensatz zur Relation `mbds_sa1_aufenthalte` in der auch weitere Kostenträger, wie zum Beispiel Länder, enthalten sind).
anz	int (1 distinkter Wert ‘1’, >55% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben
zeitraum	date (40 distinkte Werte, > 50% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben

2.4 Tabelle mbds_sa3_mel

Die Relation mbds_sa3_mel beinhaltet einem Krankenhausaufenthalt zugeordnete erbrachte medizinische Einzelleistungen, welche über die entsprechenden MEL-Codes identifiziert werden Es sind 505.719 Datensätze enthalten von 46.087 verschiedenen Patienten über die Jahre 2011 - 2016.

Attribut	Inhalt	Bemerkung
pseudonym	siehe `mbds_sa1_aufenthalte`	-
pstidtyp	siehe `mbds_sa1_aufenthalte`	-
jahr	siehe `mbds_sa1_aufenthalte`	-
ka_nr	siehe `mbds_sa1_aufenthalte`	-
mel_code	varchar(10) - Code lt. Leistungskatalog Sozialministerium (Code für	-

	ausgewählte medizinische Einzelleistungen), siehe [3].	
seite_code	varchar(2) - Körperseite auf die sich der angegebene `mel_code` bezieht.	(Inhalt: 'L','R','-',null vorhanden)
datumleistungserbr	date - Datum der Leistungserbringung. Bei Leistungen, die über einen Kalendertag hinausgehen, wird der Beginn der Leistungserbringung dokumentiert.	-
anz_leistungen	int - Anzahl der Einzelleistungen	Inhalt: numerisch von 1 bis 300
vtr	int - Versicherungsträger (siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)	
aufenthnr	siehe `mbds_sa1_aufenthalte`	-
anz2	int (1 distinkter Wert '1', >97% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben
zeitraum	date (40 distinkte Werte, > 63% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben
anz	int (1 distinkter Wert '1', >63% fehlende Werte)	unklare Bedeutung, scheint für die Auswertung und Fragestellung keine Relevanz zu haben

2.5 Tabelle vers_kate

Die Relation vers_kate beschreibt die Zugehörigkeit eines Pseudonyms zu einer bestimmten Versicherungskategorie eines Versicherungsträgers. Die einzelnen Versicherungskategorien unterscheiden sich je nach Versicherungsträger in ihren Bezeichnungen und Codesystemen. Im Datensatz sind Überschneidungen, also gleichzeitige Zugehörigkeit eines Pseudonyms zu mehreren Versicherungskategorien, vorhanden. In der Relation sind 239.178 Datensätze von 46.257 Patienten enthalten.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	-
vtr_stamm_key	int - Versicherungsträger (vgl. Kostenträger in Tabelle mbds_sa1_aufenthalte)	lt. semantischer Beschreibung nummerischer Code (=SV Code) in MBDS Daten Beschreibung

von	date - Beginn der Zugehörigkeit zu einer bestimmten Kategorie	-
bis	date - Ende der Zugehörigkeit zu einer bestimmten Kategorie	Im Datensatz enthalten sind unter Anderem Werte, die zum Zeitpunkt der Datenerfassung aktuelle Zugehörigkeiten definieren (ein Datum in der Zukunft z.B. 9998-12-31)
traeger_vkt	int - Code der Kategorie innerhalb eines Versicherungsträgers	-
trager_bez	varchar(50) - Bezeichnung der Kategorie des jeweiligen Versicherungsträgers	-

2.6 Tabelle aufenth_kette

Die Relation aufenth_kette beschreibt den Zusammenhang mehrerer Krankenhausaufenthalte (Relation mbds_sa1_aufenthalte). Der definierte Zeitraum in dieser Relation soll mehrere Krankenhausaufenthalte zusammenführen um z.B. Verlegungen, Zuweisungen, neuerliche Einweisungen zu berücksichtigen. Es sind 197.086 Datensätze von 47.166 Patienten aus den Jahren 2010 bis 2016 vorhanden.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	-
aufenthkettennr	int - Fortlaufende Nummer der Kette	Identifiziert zusammen mit dem Attribut pseudonym eindeutig eine Aufenthaltskette (Fortlaufende Nummer pro Patient).
aufenthkette_begin	date - Beginn einer Aufnahmekette	-
aufenthkette_ende	date - Ende einer Aufnahmekette	-
alterbeientlass	int - Alter am Tag der Entlassung (`aufenthkette_ende`)	Eine Prüfung der Daten ergibt abweichende Werte beim Attribut `alterbeientlass` wenn dieses Alter aus dem Attribut `aufenthkette_ende` und dem Attribut `gebdatum` aus der Tabelle `stammdaten` berechnet wird
geschl	varchar(2) - Geschlecht des Patienten	Textuell kodiert als 'M' bzw. 'W'. Eine Prüfung ergibt teilweise inkonsistenten Datenbestand; die Geschlechter der `aufenth_kette` stimmen nicht immer mit den Geschlechtern der Tabelle `mbds_sa1_aufenthalte` überein.

sterbefall	varchar(2) - Ende der Aufenthaltskette durch Sterbefall	Textuell kodiert als 'J' bzw. 'N'.

2.7 Tabelle ek

In der Relation ek werden Leistungen aus dem niedergelassenen Bereich abgebildet. Es sind 18.049.786 Daten vorhanden von 46.074 unterschiedlichen Pseudonymen aus den Jahren 2011 bis 2016.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	-
ltim	date -Datum der Leistungserbringung	Wertebereich 2008-02-13 - 2017-03-30
posnr	varchar(50) - Nummer der Leistungsart.	Wird näher beschrieben durch das Attribut `posbez`.
anzahl	varchar(50) - Anzahl der erbrachten Leistungen	-
betrag	varchar(50) - Betrag in Euro	-
fgr	int - Code der Fachgruppe des Leistungserbringers	vgl. Tabelle Fachgebiet in GAP-DRG
posbez	Text - textuelle Beschreibung der Leistung	Die textuelle Beschreibung ist häufig nicht ident mit der Codierung im Attribut `posnr`, gegebenenfalls ist eine Überarbeitung notwendig.
kat	text - Kategorie	-
honoid	int	unklare Bedeutung, es scheint sich aber über die Zuordnung von Honoraren zu handeln; der Eintrag ist in 80% der Fälle null
vtr	int - Versicherungsträger	(siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`)
jahr	int - Jahr	Jahr in dem die Abrechnung durchgeführt wurde
quartal	int - Quartal	Quartal der Abrechnung

2.8 Tabelle au

Die Relation au beschreibt die Arbeitsunfähigkeit einer Person, welche über einen gewissen Zeitraum mit einer Diagnose dokumentiert wurde. Es sind 100.288 Datensätze zu 12.898 verschiedenen Patienten aus den Jahren 2011 bis 2016 vorhanden.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	-

dauer_gesamt	int - Dauer der Arbeitsunfähigkeit	-
beginn	date - Beginndatum der Arbeitsunfähigkeit	teilweise unrealistische Werte enthalten (z.B. 2999-12-31)
bisdatum	date - Enddatum der Arbeitsunfähigkeit	teilweise unrealistische Werte enthalten (z.B. 3000-01-01)
diag	varchar(50) - Diagnosecode lt. ICD	Im den meisten Datensätzen wird ICD-10 als Diagnosecodierung verwendet. In den Daten befinden sich allerdings auch 36 verschiedene ICD-9 Codes.
vtr	int - Versicherungsträger	siehe Kostenträger in Tabelle `mbds_sa1_aufenthalte`
jahr	int - Jahr	Jahr des Endes der Arbeitsunfähigkeit (bisdatum)

2.9 Tabelle hm

In der Relation hm werden die bezogenen Heilmittel (Medikamente, etc.) zu einem Patienten verzeichnet. Im Datensatz sind 9.419.633 Einträge zu 46.072 Patienten aus den Jahren 2011 - 2016 vorhanden.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	
ltim	date - Datum der Leistungserbringung	
summe_anzahl	varchar(50) - Anzahl der Heilmittel	
summe_betrag	varchar(50) - Betrag der Heilmittel (Kosten in Euro)	
pharmanr	varchar(50) - Pharmazentralnummer	siehe Apothekerkammer[1] , teilweise fehlende Werte (143, mit 'XXXXXXXX' codiert)
atc	varchar(50) - ATC Code	Teilweise Fehlende Werte, diese Werte sollten allerdings aus der Pharmazentralnummer generiert werden können

2.10 Tabelle reha

Zusätzlich zu den bisher behandelten Relationen wurden im späten Verlauf des Projekts Daten zu Aufenthalten in Rehakliniken erhalten. Die beschriebenen Attribute sind die für die Auswertung relevanten.

[1]https://www.apotheker.or.at/Internet/OEAK/NewsPresse_1_0_0a.nsf/agentEmergency!OpenAgent&p=6213BA470504C468C1256E8900370485&fsn=fsStartHomeFachinfo&iif=0

In der Relation reha werden die abgeschlossenen Rehabilitationsaufenthalte zu einem Patienten verzeichnet. Im Datensatz sind 7.494 Einträge zu 5.696 Patienten aus den Jahren 2011 - 2016 vorhanden.

Attribut	Inhalt	Bemerkung
pseudonym	varchar(50) - Patientenkennung	
diag1	varchar(10) - Hauptdiagnose	ICD-10 Codes im Format `X00` – `X0000`
beginn	varchar(8) – Beginn des Aufenthaltes	(Format `DDMMYYYY`)
ende	varchar(8) - Ende des Aufenthaltes	(Format `DDMMYYYY`)
aufart	varchar(1) - Aufnahmeart	Enthält ausschließlich den Wert `R` (Aufnahme in den Bereich der Rehabilitation)
entart	varchar(1) - Entlassungsart	Enthält Werte `E`, `T`, `2`, `S` Siehe Tabelle mbds_sa1_aufenthalte -> entlassungsart

Die Bedeutung des Wertes `2` im Attribut `entart` ist unbekannt, jedoch ist die Art der Entlassung angesichts der durchgeführten Auswertung nebensächlich.

3 Datenbankschema

Die im obigen Abschnitt beschriebenen Rohdaten werden in ein dem Projektumfang und der Projektfragestellung entsprechendes Schema überführt. Die dabei getätigten Annahmen und Arbeitsschritte sowie die Struktur des neuen Schemas werden im Folgenden festgehalten. Weiters sind zu den einzelnen Relationen einfache statistische Kennwerte angegeben, betreffend den darin enthaltenen Daten. Eine vollständige graphische Darstellung des Datenbankschemas ist im Anhang zu finden (siehe Abbildung 2).

3.1 Inhalt

Das vorliegende Datenbankschema enthält im Wesentlichen die bereinigten Daten der Rohdaten mit zusätzlich eingefügtem Material zur Annotation der Daten und Material, welches aus technischer Sicht erforderlich ist für die Datenbehandlung.

Überblicksmäßig sind Daten zu Versicherten aus den Jahren 2011-2016 vorhanden. Diese Daten sind bereits vorselektiert von Seiten der Auftraggeber im Hinblick auf die Analyse von Herzinfarktpatienten und bestehen aus folgenden wesentlichen Teilen.

- Aufenthalte: Spitalsaufenthalte samt zugeordneter Diagnosen, medizinischen Einzelleistungen sowie weiterer Informationen (Dauer der Aufenthalte, Verlegungen, Tod). Die Diagnosen werden als ICD-10 BMGF 2017 [4], und die medizinischen Einzelleistungen laut Leistungskatalog 2018 des BMGF [3] codiert
- Heilmittel: Über Krankenversicherungen abgerechnete Heilbehelfe und Medikamente samt ihrer Stoffgruppe in Form ihrer ATC-Codierung (Anatomical Therapeutic Chemical-Codes der WHO) in ihrer aktuellen Fassung.
- Leistungen aus dem niedergelassenen Bereich: Über die Krankenversicherungen abgerechnete Leistungen im Rahmen der niedergelassenen Versorgung. Hierzu sind die Arten der durchgeführten Leistungen, Anzahl und Zeitpunkte vorhanden.
- Patienten: Stammdaten der Patienten enthalten neben Geschlecht und Alter auch ein eventuelles Sterbedatum und ihren Wohnort sowie die Zugehörigkeit zu einer Versicherung im zeitlichen Verlauf.

Eine genauere Beschreibung der einzelnen Relationen erfolgt in den folgenden Kapiteln.

3.2 Namenskonvention

Die Tabellen werden, zusätzlich zu ihrer Benennung entsprechend dem Inhalt, nach der jeweiligen Datenquelle mit einem Präfix benannt. Alle Rohdaten stammen aus der LEICON Datenbank und werden entsprechend mit dem Präfix `lei_` versehen. Weitere Daten, die zum Zwecke der Annotation der bestehenden Daten eingefügt werden, werden mit dem Präfix `mat_` (für Material) versehen. Dazu zählen beispielsweise ICD-10 Codes, etc. Des Weiteren werden Attributnamen möglichst sprechend und einheitlich benannt. Daraus ergibt sich teilweise der Bedarf bestehende Attributnamen zu verändern.

Präfix	Bedeutung
`lei_`	Tabellen welche aus den Rohdaten und daher aus der LEICON Datenbank stammen.
`mat_`	Zusätzliche Tabellen zum Zwecke der Annotation bestehender Daten. Hier werden beispielsweise Daten aus bestehenden Code Systemen eingefügt.
`work_`	Tabellen die für den Arbeitsablauf nötig erscheinen werden mit diesem Präfix versehen. Beispielsweise wird die Entscheidung ob ein Individuum aus der Forschungskohorte ausgeschlossen wird in der Tabelle `work_kohorte_ausschluss` gespeichert.

Tabellen und Attribute werden in deutscher Sprache bezeichnet. Bei der Vergabe von Bezeichnungen wird auf eine semantische Gruppierung von mehreren Attributen geachtet.

In entsprechenden Abfragen wird darauf geachtet Attribute möglichst eindeutig zu beschreiben. So wird auf das Geburtsdatum des Individuums beispielsweise mit `lei_stammdaten.pat_geburt` zugegriffen anstatt nur den Attributnamen zu verwenden. Attribute welche eine nachträglich eingefügte laufende Nummer enthalten werden mit dem Präfix `lfd_` (für laufend) benannt.

Die DDL (Data Definition Language) SQL Artefakte zum Anlegen der im Folgenden beschriebenen Relationen sind im Projektrepository am GitLab Server des DEXHELPP Systems unter dem Pfad `qi-ka/schema_modify/*` zu finden. Ebenso sind darin weitere Artefakte zum Manipulieren der Daten enthalten. Alle Skripts werden mit nummerischen Präfixen versehen um die Ausführreihenfolge festzulegen. Diese Dateien werden nach der fortlaufenden Nummer mit dem Namen der Tabelle auf die sich diese hauptsächlich beziehen bezeichnet und folgend eine Bezeichnung für die ausgeführte Operation.

3.3 Darstellung des Datenbankschemas

In Abbildung 2 wird eine Komplettansicht des Datenbankschemas gezeigt. Für die drei Datenbereiche LEICON-Tabellen (lei_), Material-Tabellen (mat_) und Arbeits-Tabellen (work_) werden in den nachfolgenden Kapiteln auch Abbildungen der Detail-Schemas gezeigt.

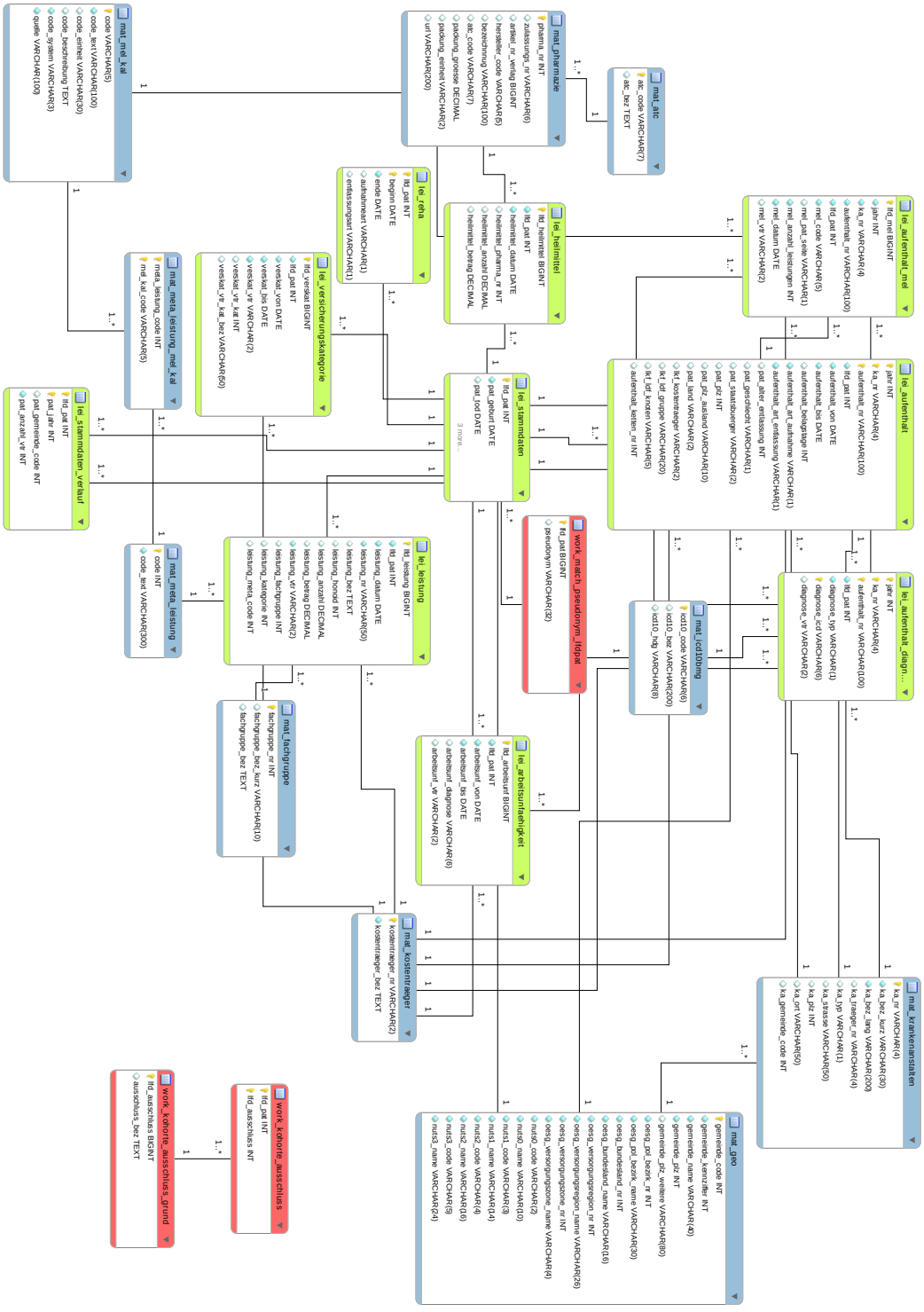

Abbildung 2: Darstellung des Datenbankschemas mit allen Relationen
Aus den Rohdaten (blau), zusätzliche annotierende Relationen (grün) und für die Projektabwicklung notwendige Relationen (rot)

3.4 Zusammenfassung

Die Rohdaten werden mit entsprechenden Anpassungen und Bereinigungen in das neue Schema überführt. Eine Relation (`work_match_pseudonym_lfdpat`) stellt einen Bezug zwischen den in den Rohdaten vorhanden Pseudonymen (als Bezeichnung für einen Patienten) und den neu eingeführten Patientenlaufnummern (`lfd_pat`) her. Diese wird eingeführt um eine leichtere Lesbarkeit zu gewährleisten und um die Pseudonyme bei Auswertungen nicht aus der Datenbank exportieren zu müssen. Zwei weitere Relationen definieren den Ausschluss aus der Forschungskohorte (`work_kohorte_ausschluss` & `work_kohorte_ausschluss_grund`) wobei die Begründung für einen Ausschluss jeweils vermerkt wird. So werden beispielsweise unplausible bzw. als fehlerhaft erachtete Datensätze ausgeschlossen.

Im Folgenden werden die Struktur und der Inhalt der einzelnen Relationen beschrieben. Die einzelnen enthaltenen Attribute werden in diesem Dokument als Tabellen dargestellt wobei der Name, der Datentyp (entsprechend der PostgreSql Datentypen), ein Verweis auf dessen Einschränkungen (PK = Primary Key, FK = Foreign Key, NN = Not Null) und eine semantische Erklärung angegeben wird. Zusätzlich wird der Prozess der Datenüberführung und eventuelle Ausschlüsse aus der Kohorte beschrieben. Ebenso werden überblicksmäßige Statistiken über die enthaltenen Daten angeführt.

3.5 Leicon-Tabellen (`lei_`)

Die folgenden Tabellen entsprechen im Wesentlichen den Tabellen der Rohdaten in aufbereiteter Form. Zum Zwecke der einfacheren Lesbarkeit und Einheitlichkeit werden die Tabellennamen sowie Attributnamen angepasst.

Wie in den vorhergehenden Abschnitten wird die Struktur der Tabellen beschrieben, aber zusätzlich wird in den folgenden Abschnitten der Ablauf der Datenmigration detaillierter beschrieben. Die beschriebenen Relationen sind überblicksmäßig in Abbildung 3 dargestellt.

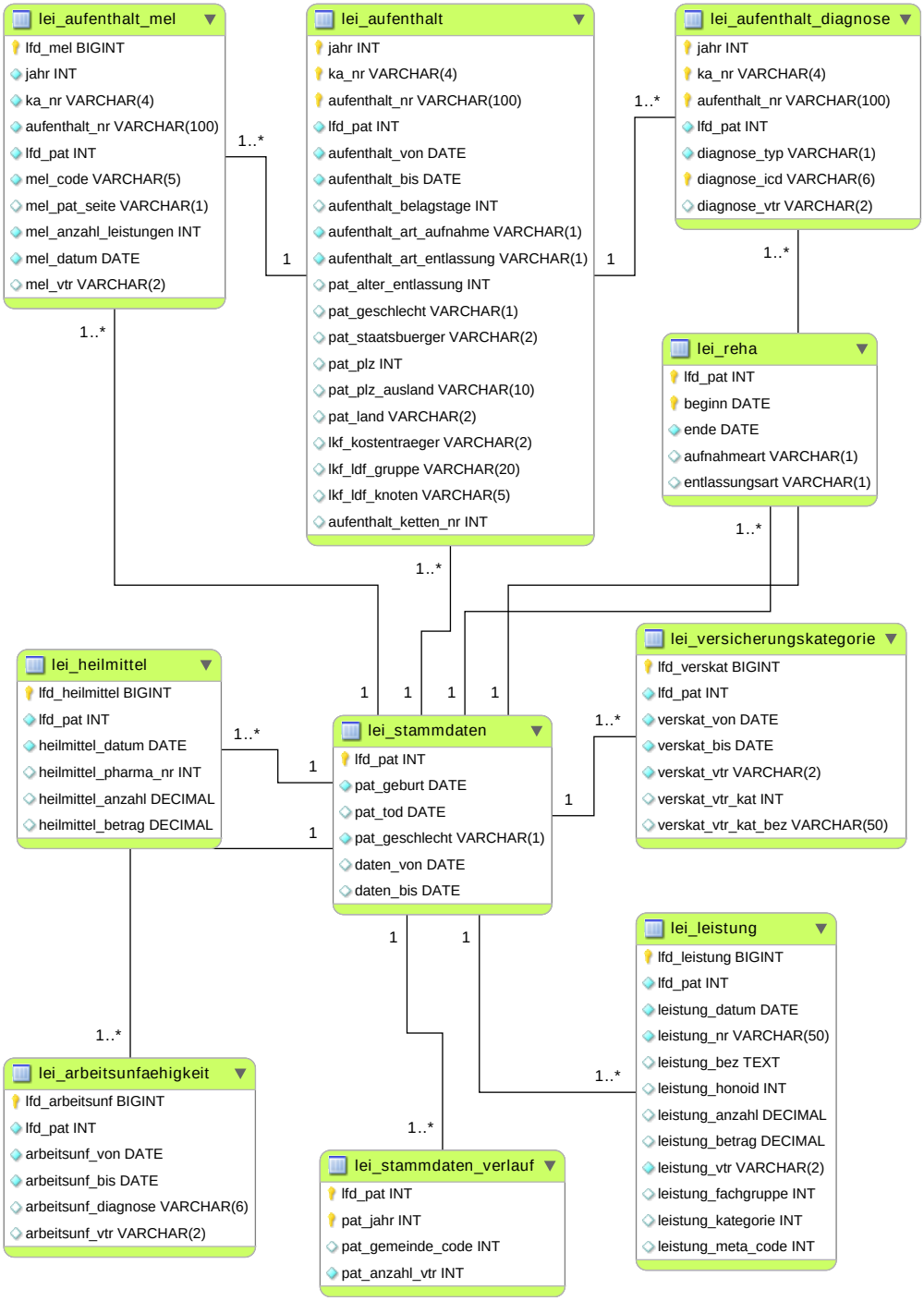

*Abbildung 3: Überblick über die Relationen mit dem Präfix lei_*

3.5.1 Tabelle `lei_stammdaten`

Einem Individuum zugeordnete unveränderliche Eigenschaften wie z.B. das Geschlecht, Geburtsdatum- und Todesdaten.

Tabelle:	`lei_stammdaten`
Beschreibung:	Stammdaten der Individuen aus der Leicon Datenbank

Anzahl Einträge:	46.231
PK:	`lfd_pat`
FK:	`lfd_pat` -> `work_match_pseudonym_lfdpat`
Anmerkung:	entspricht teilweise der Tabelle `stammdaten` in Rohdaten

SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/110_lei_stammdaten_create.sql`
	`DEXHELPP-GITLAB: qi-ka/schema_modify/111_lei_stammdaten_transfer.sql`
	`DEXHELPP-GITLAB: qi-ka/schema_modify/200_lei_stammdaten_insert_latest_data.sql`

Die Stammdaten der Individuen sind in dieser Tabelle enthalten und stammen aus dem Leicon Datensatz. Die überführten Daten entsprechen im Wesentlichen denen der Rohdaten. Es existieren Daten von Patienten zu mehreren Jahren. Ein dem Patient zugeordnetes Pseudonym hat in den Rohdaten daher mehrere Einträge. Da die Relation einige zeitlich nicht veränderliche Attribute beinhaltet, werden die Attribute der Relation welche einer zeitlichen Veränderung unterworfen sind in eine eigene Relation ausgelagert. Dadurch wären die Datensätze in dieser Tabelle eindeutig über das Attribut `lfd_pat` identifizierbar und könnten als Referenz in anderen Relationen verwendet werden. Wie in einem vorhergehenden Abschnitt beschrieben werden alle vorkommenden Pseudonyme entsprechend in die Tabelle `work_match_pseudonym_lfdpat` übernommen und danach die Laufnummer `lfd_pat` zur eindeutigen Identifizierung der Datensätze verwendet.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
`lfd_pat`	Integer	PK, FK	Eindeutige Kennung des Individuums
`pat_geburt`	Date	NN	Geburtsdatum des Individuums
`pat_tod`	Date	-	Sterbedatum des Individuums
`pat_geschlecht`	Varchar(1)	NN, IN('M', 'W')	Geschlecht des Individuums
`daten_von`	Date	-	Datum der frühesten verfügbaren Daten
`daten_bis`	Date	-	Datum am ältesten verfügbaren Daten

Überführen der Rohdaten:

Die Daten stammen aus der Relation `import.stammdaten` der Rohdaten welche in diese und in die nachfolgend beschriebene Relation (`lei_stammdaten_verlauf`) aufgespalten werden. Die Pseudonyme werden in die entsprechende Relation überführt. Im folgenden Arbeitsablauf werden inkonsistente Datensätze (im Besonderen `pat_geburt`, `pat_tod` und `pat_geschlecht`) aus der Kohorte über entsprechende Verweise in den Relationen `work_kohorte_ausschluss` und `work_kohorte_ausschluss_grund` ausgeschlossen. Durch diesen Mechanismus werden aus den ursprünglich 255.505 Datensätzen in der Relation `stammdaten` (welche Patienten pro Jahr beinhalten) 4 Patienten mit fehlerhaften

Daten ausgeschlossen, was 22 Datensätzen in der ursprünglichen Relation entspricht. Die Daten von 15 Pseudonymen werden manuell nachbearbeitet, da sich beispielsweise nur ein Geburtsdatum von den übrigen eingetragenen Geburtsdaten unterscheidet. Diese Irregularitäten betreffen gesamt 9 Datensätze und treten nur bei Patienten ab einem Alter von 100 Jahren auf. Diese Fehler werden manuell nachbearbeitet (eine zweistellige Eingabe für das Geburtsjahr könnte eine Erklärung für diese Inkonsistenzen sein). Entsprechend wird mit alternierenden Geschlechtern umgegangen. Bei einem Datensatz wird ein fehlendes Geschlecht aufgrund anderer Datensätzen ergänzt. Bei fünf Datensätzen werden Geschlechter geändert welche in den übrigen (>75%) Datensätzen anders vorliegen. Es verbleibt ein unschlüssiger Fall welcher in weiterer Folge ausgeschlossen wird.

Ebenso wird mit wechselnden Todesdaten verfahren (betrifft lediglich drei Datensätze), diese Datensätze werden aus der Forschungspopulation ausgeschlossen.

Um die Erstellung der Auswertungen zu erleichtern werden zusätzlich die Attribute `daten_von` und `daten_bis` befüllt. Diese stellen zum einen das früheste Datum einer Aufzeichnung jeglicher Art zu einem Patienten dar und zum Anderen das Datum der letzten Aufzeichnung. Es werden alle Daten aus den Relationen lei_arbeitsunfaehigkeit (arbeitsunf_von, arbeitsunf_bis), lei_aufenthalt (aufenthalt_von, aufenthalt_bis), lei_heilmittel (heilmittel_datum), lei_leistung (leistung_datum), lei_versicherungskategorie (verskat_von, verskat_bis - wenn nicht als Wert 9998-12-31 vorhanden) betrachtet.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 4 veranschaulicht.

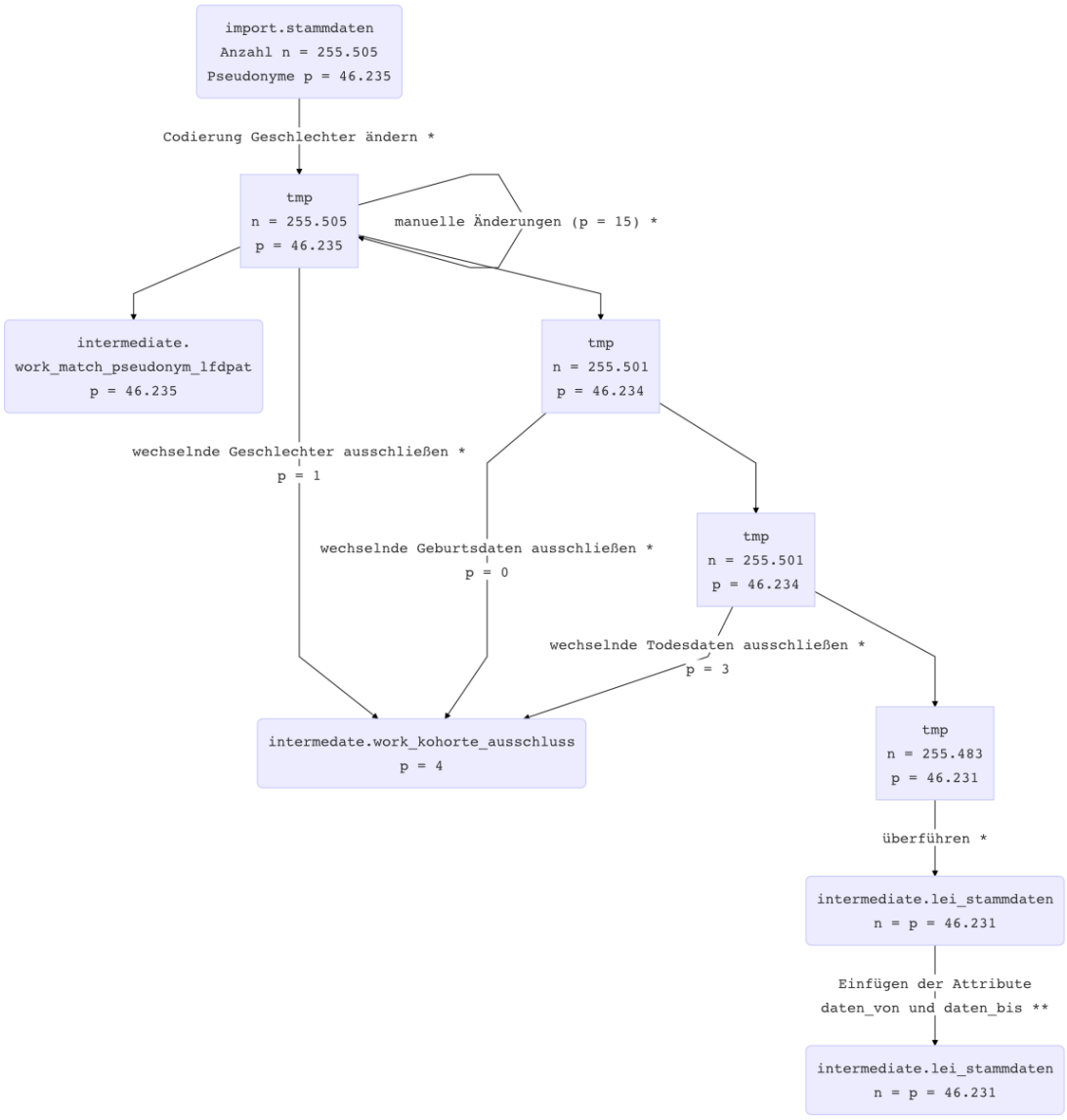

Abbildung 4: Überführung der Stammdaten

Skript 111_lei_stammdaten_transfer.sql (*) und 200_lei_stammdaten_insert_latest_data.sql (**)

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_stammdaten
    ORDER BY RANDOM() LIMIT 5;
```

lfd_pat	pat_geburt	pat_tod	pat_geschlecht	daten_von	daten_bis
27979	1952-07-25	-	M	2011-10-01	2016-12-23
32607	1936-07-22	2016-07-05	M	2011-01-01	2016-07-05
20203	1936-02-10	-	M	2011-01-03	2016-12-27
11828	1933-01-31	-	W	2011-01-27	2016-12-21
3611	1931-12-25	-	W	2011-01-24	2016-12-20

```
/* there is only 1 entry per unique patient */
SELECT COUNT(*) AS stammdaten, COUNT(DISTINCT(lfd_pat)) AS distinct_lfd_pa
t
    FROM intermediate.lei_stammdaten;
```

stammdaten	distinct_lfd_pat
46231	46231

```
/* number of males/females */
SELECT pat_geschlecht, COUNT(*)
    FROM intermediate.lei_stammdaten
    GROUP BY pat_geschlecht;
```

pat_geschlecht	count
M	29679
W	16552

3.5.2 Tabelle `lei_stammdaten_verlauf`

Abbildung zeitlich veränderlicher Eigenschaften der Stammdaten eines Individuums.

Tabelle:	`lei_stammdaten_verlauf`
Beschreibung:	Beinhaltet zeitlich veränderliche Daten der Stammdaten
Anzahl Einträge:	255.483
PK:	`lfd_pat`, `pat_jahr`
FK:	`lfd_pat` -> `lei_stammdaten`, `pat_gemeinde_code` -> `mat_geo`
Anmerkung:	-

SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/120_lei_stammdaten_verlauf_create.sql`
	`DEXHELPP-GITLAB: qi-ka/schema_modify/121_lei_stammdaten_verlauf_transfer.sql`
	`DEXHELPP-GITLAB: qi-ka/schema_modify/122_lei_stammdaten_verlauf_cleanGeoCode.sql`
	`DEXHELPP-GITLAB: qi-ka/schema_modify/123_lei_stammdaten_verlauf_fk.sql`

Diese Relation erweitert die Relation `lei_stammdaten` um zeitlich veränderliche Attribute und wird daher über die Laufnummer der Patienten und das Jahr identifiziert.

Attribut	Typ	Verweis	Bemerkung
`lfd_pat`	Integer	PK, FK	Eindeutige Kennung des Individuums
`pat_jahr`	Integer	PK	Jahr auf den sich dieser Datensatz bezieht
`pat_gemeinde_code`	Integer	FK	Gemeindecode (GCD) lt. Adressregisterverordnung, siehe Statistik Austria [2]
`pat_anzahl_vtr`	Integer	NN	Anzahl der zugeordneten Versicherungsträger in diesem Jahr

Überführen der Rohdaten:

Die Daten stammen aus der Relation `import.stammdaten` der Rohdaten welche in die zuvor beschriebene Relation (`lei_stammdaten`) und der vorliegenden Relation (`lei_stammdaten_verlauf`) aufgespalten werden. Hier werden die zeitlich veränderlichen Informationen zu einem Stammdatensatz gespeichert. Bereits in den vorherigen Schritten ausgeschlossene Individuen werden nicht behandelt und von der Bearbeitung ausgeschlossen.

Die in den Rohdaten vorkommenden Attribute *gemnr* und *geo_key*, werden aufgrund inkonsistenter Daten zusammengeführt in das Attribut `gemeinde_code`. Die beiden Attribute enthalten Daten zu den Wohnorten der Patienten und unterscheiden sich nur in der Unterteilung der Bezirke in Wien (vgl. Gemeindecode und Gemeindekennziffer lt. Statistik Austria [2]). Die Daten in den beiden Attributen lassen eine falsche Eintragung von Werten vermuten. Beispielsweise die Gemeindekennziffer als Gemeindecode verwendet und umgekehrt. Es wird in weiterer Folge jeweils die Ausprägung mit der höheren Genauigkeit verwendet um den Wohnort des Patienten zu beschreiben.

Danach werden die gemeinde_codes entsprechend der verschiedenen Änderungen (z.B. Gemeindezusammenlegungen) auf den aktuellen Gebietsstand 2018 lt. Statistik Austria gebracht. Dazu werden die amtlichen Verkündungen dieser Gebietsänderungen auf unsere Daten angewendet. Der Prozess kann in der Datei ‘data/regionen/anpassungen_lei_stammdaten_verlauf.xls’ nachverfolgt werden. Diese Änderungen werden einem eigenständigen SQL Script (‘122_lei_stammdaten_verlauf_cleanGeoCode.sql’) auf die Daten angewendet.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 5 veranschaulicht.

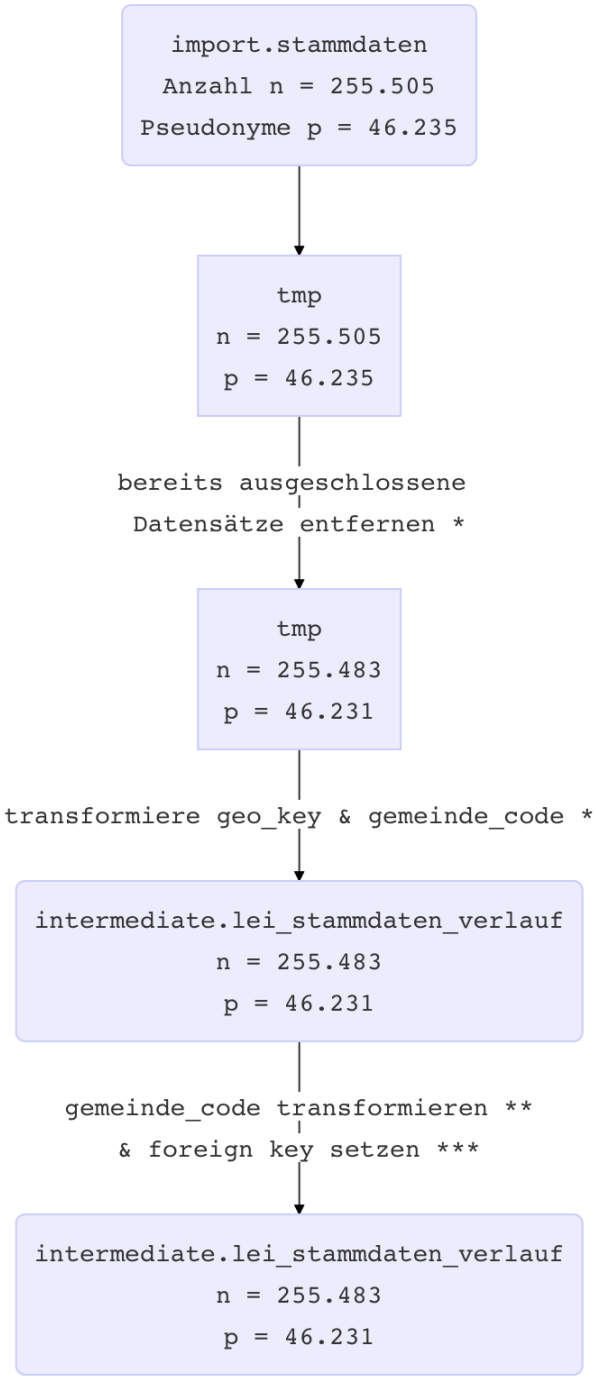

Abbildung 5: Überführung der zeitlich veränderlichen Daten zu den Stammdaten Skripte 121_lei_stammdaten_verlauf_transfer.sql (*), 122_lei_stammdaten_verlauf_cleanGeoCode.sql (**) und 123_lei_stammdaten_verlauf_fk.sql (***)

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_stammdaten_verlauf
    ORDER BY RANDOM() LIMIT 5;
```

lfd_pat	*pat_jahr*	pat_gemeinde_code	pat_anzahl_vtr
27385	2016	40401	1

33104	2016	20421	1
2069	2016	40101	1
29093	2015	70528	2
39672	2014	80303	1

```
/* distinct_patient_count vs. row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_stammdaten_verlauf;
```

distinct_patient_count	row_count
46231	255483

```
/* no. of insurances per patient per year */
SELECT pat_anzahl_vtr, COUNT(pat_anzahl_vtr)
    FROM intermediate.lei_stammdaten_verlauf
    GROUP BY pat_anzahl_vtr
    ORDER BY COUNT(pat_anzahl_vtr) DESC;
```

pat_anzahl_vtr	count
1	124800
2	84890
3	30082
4	13106
5	1533
6	760
7	252
8	51
9	7
10	2

3.5.3 Tabelle `lei_versicherungskategorie`

Bildet die Zugehörigkeit von Individuen zu einer Versicherungskategorie eines Versicherungsträgers und deren zeitliche Veränderung ab. Es sind 239.102 Einträge von 46.231 verschiedenen Patienten vorhanden.

Tabelle:	`lei_versicherungskategorie`
Beschreibung:	Zugehörigkeit eines Individuums zu einer bestimmten Versicherungskategorie eines Versicherungsträgers
Anzahl Einträge:	239.102
PK:	`lfd_verskat`
FK:	`lfd_pat` -> `lei_stammdaten`, `verskat_vtr` -> `mat_kostentraeger`
Anmerkung:	entspricht Tabelle `vers_kate` in Rohdaten
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/130_lei_vers_kategorie_create.sql`

DEXHELPP-GITLAB:
qi-ka/schema_modify/131_lei_vers_kategorie_transfer.sql

Die Zugehörigkeit eines Individuums zu einer Versicherungskategorie in einem Zeitraum wird durch diese Tabelle beschrieben. Auf eine weitere Bearbeitung der Daten wird hinsichtlich der Relevanz dieser Daten im Bezug auf die Aufgabenstellung des Projekts vorerst verzichtet.

Attribut	Typ	Verweis	Bemerkung
lfd_vers_kat	Serial	PK	Identifizierung eines Datensatzes
lfd_pat	Integer	NN, FK	Eindeutige Kennung des Individuums
verskat_von	Date	NN	Startzeitpunkt der Zugehörigkeit
verskat_bis	Date	NN	Endzeitpunkt der Zugehörigkeit
verskat_vtr	Varchar(2)	NN, FK	Versicherungsträger dem dieser Patient im Zeitraum zugeordnet ist
verskat_vtr_kat	Integer	-	Versicherten-Kategorie (des jeweiligen Trägers)
verskat_vtr_kat_bez	Varchar(50)	-	Bezeichnung der Kategorie (des jeweiligen Trägers)

Überführen der Rohdaten:

Das Überführen der Daten gestaltet sich aufgrund der Ähnlichkeit der Tabelle zur Ursprünglichen sehr einfach. Im Rohdatensatz sind Daten zu Individuen enthalten welche in der Relation lei_stammdaten nicht vorkommen. Diese werden über entsprechende Verweise ausgeschlossen. Hierzu wird in der Tabelle work_kohorte_ausschluss_grund eine Begründung eingetragen und in weiterer Folge Pseudonyme mittels Eintrag in work_kohorte_ausschluss ausgeschlossen. Zusätzlich werden bisher nicht bekannte Pseudonyme in der Tabelle work_match_pseudonym_lfdpat vermerkt.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 6 veranschaulicht.

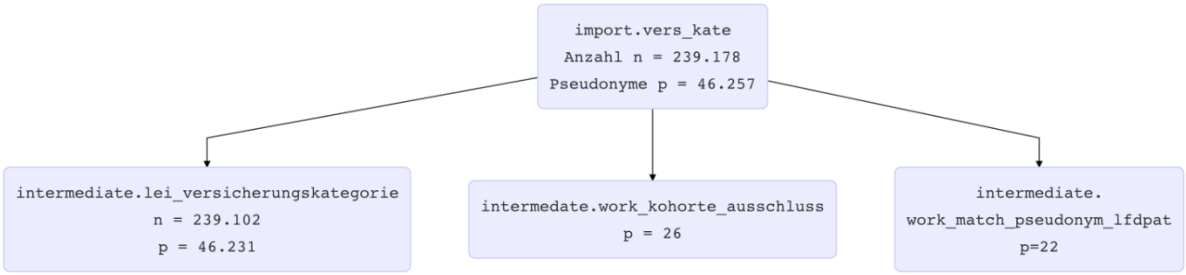

Abbildung 6: Überführung der Daten zu Versicherungskategorien
Skript 131_lei_versicherungskategorie_transfer.sql

```
/* 5 random entries */
SELECT *
```

```
    FROM intermediate.lei_versicherungskategorie
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

lfd_verskat	*lfd_pat*	verskat_von	verskat_bis	verskat_vtr
119236	985	2010-07-01	2010-12-31	11
80461	38005	2002-01-01	9998-12-31	19
77777	35434	2012-04-01	2012-06-30	11
130473	1963	2008-01-01	2012-12-31	11
9911	2553	2009-09-01	2009-10-31	07

verskat_vtr_kat	verskat_vtr_kat_bez
2	Angestellter
8	Sonstiger
1	Arbeiter
8	Sonstiger
0	-

```
/* distinct_patient_count vs. row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_versicherungskategorie;
```

distinct_patient_count	row_count
46231	239102

```
/* no. of distinct categories */
SELECT COUNT(DISTINCT verskat_vtr_kat)
    FROM intermediate.lei_versicherungskategorie;
```

count
43

```
/* earliest beginning, latest ending */
SELECT MIN(verskat_von) AS earliest_beginning, MAX(verskat_bis) AS latest_
ending
    FROM intermediate.lei_versicherungskategorie;
```

earliest_beginning	latest_ending
1998-01-01	9998-12-31

3.5.4 Tabelle `lei_aufenthalt`

Enthält Spitalsaufenthalte der beobachteten Individuen. Diesen Aufenthalten werden in weiterer Folge Diagnosen und medizinische Einzelleistungen zugeordnet. Aufenthalte beschreiben klar unterscheidbare Aufnahmen in Krankenanstalten. Im vorliegenden Datensatz sind 239.907 Aufenthalte von 46.231 verschiedenen Patienten enthalten. Diese Relation steht in enger Verbindung zu den ihr zugeordneten Diagnosen und medizinischen Einzelleistungen.

Tabelle:	lei_aufenthalt
Beschreibung:	Beinhaltet Datensätze zu Spitalsaufenthalten
Anzahl Einträge:	239.907
PK:	jahr, ka_nr, aufenthnr
FK:	ka_nr -> mat_krankenanstalten, lfd_pat -> lei_stammdaten, lkf_kostentraeger -> mat_kostentraeger
Anmerkung:	entspricht Tabelle mbds_sa1_aufenthalte in Rohdaten
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/140_lei_aufenthalt_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/141_lei_aufenthalt_transfer.sql DEXHELPP-GITLAB: qi-ka/schema_modify/142_lei_aufenthalt_cleanPlz.sql

Wie unten beschrieben wurden einige Attribute betreffend der LKF Punkteberechnung aus dem Datensatz entfernt um die Relation übersichtlicher zu gestalten. Da sich diese Attribute auf die Finanzierung des Gesundheitswesens beziehen, und diese in der vorliegenden Studie nicht betrachtet wird, kann die Entfernung ohne Konsequenzen für unsere weitere Arbeit erfolgen.

Die Relation import.aufenth_kette der Rohdaten beschreibt zusammenhängende Spitalsaufenthalte in einem gewissen Zeitraum, beispielsweise bei Verlegung des Patienten. Diese Relation wird nicht in dieses Schema überführt, sondern in die vorliegende Relation (lei_aufenthalt) integriert.

Attribut	Typ	Verweis	Bemerkung
jahr	Int	PK	Jahr des Aufenthaltes
ka_nr	Varchar(4)	PK, FK	Kennung des jeweiligen Krankenhauses
aufenthalt_nr	Varchar(100)	PK	Innerhalb eines Jahres und Krankenhauses eindeutige Nummer eines Aufenthaltes
lfd_pat	Integer	NN, FK	Eindeutige Kennung des Individuums
aufenthalt_von	Date	NN	Aufnahmedatum
aufenthalt_bis	Date	NN	Entlassungsdatum
aufenthalt_belagstage	Integer	-	Summe der Mitternachtsstände
aufenthalt_art_aufnahme	Varchar(1)	NN	Leistungsbereich der Aufnahme (siehe GAP-DRG semantisches Datenbankmodell - MBDS Daten Beschreibung [1]). Zusätzlich zu den Beschreibungen der Werte

			im semantischen Datenmodell (siehe oben) kommt der Wert 'K' mit der Bedeutung 'Aufnahme eines 0-Tagesfalls' hinzu.
aufenthalt_art_entlassung	Varchar(1)	NN	Art der Entlassung (siehe GAP-DRG) zusätzlich Wert '4' - Entlassung gegen Revers (lt. Rückfrage mit NÖGKK)
pat_alter_entlassung	Integer	-	Alter des Individuums bei Entlassung
pat_geschlecht	Varchar(1)	-	Geschlecht des Individuums (m/w)
pat_staatsbuerger	Varchar(2)	-	Staatsbürgerschaft des Individuums; Ländercode lt. Sozialministerium
pat_plz	Int	-	Hauptwohnsitz des Individuums (österreichische PLZ oder Wert "AUSL")
pat_plz_ausland	Varchar(10)	-	siehe Hauptwohnsitz, bei Hauptwohnsitz im Ausland
pat_land	Varchar(2)	-	Wohnsitzland im Ausland (optional), Ländercodes lt. Sozialministerium
lkf_kostentraeger	Varchar(2)	FK	Kostenträgercode
lkf_ldf_gruppe	Varchar(20)	-	Fallpauschalengruppe lt. LKF (= Leistungsorientierten Diagnosenfallgruppen)
lkf_ldf_knoten	Varchar(5)	-	Fallpauschale (A-F) falls reguläre Fallgruppe eingetragen (lt. Dokument)
aufenthalt_ketten_nr	Integer	-	Beschreibt zusammengehörige Aufenthalte. Die Werte sind generiert, jedoch werden Zusammenhänge aus `import.aufenth_kette` damit beschrieben

Das Attribut `aufenthalt_art_aufnahme` enthält neben den dokumentieren Codes entsprechend der GAP-DRG2 [1] einen zusätzlichen Code (Wert: K). Die Bedeutung des Wertes wird anhand von Daten aus dem Projekt ADE-PIM vervollständigt. Der Code 'K' entspricht einer 'Aufnahme eines 0-Tagesfalls'

Das Attribut aufenthalt_art_entlassung enthält neben den durch die GAP-DRG2 [1] erklärbaren Codes zusätzlich einen Wert '4'. Welchem die Bedeutung 'Entlassung gegen Revers' zugeschrieben wird. Diese Erkenntnis ist gestützt auf Daten der NÖGKK.

Die Attribute welche sich mit der Punkteberechnung lt. LKF befassen werden vorerst ausgeschlossen, (dabei handelt es sich um die Attribute punkteldftageskomp, punkteldfleistungskomp, ptbeldauausruntentag, ptbeldauausruntenleistung, zusatzptbeldauausroben, zusatzpt_intensiv, zusatzpt_mehrfachleist, pt_spez_bereich, pt_total). Des Weiteren wurden die beiden Attribute zeitraum und anz ausgeschlossen, denen keine nachvollziehbare Bedeutung zugeordnet werden konnte.

Überführen der Rohdaten:

Die Daten werden ähnlich den ursprünglichen Daten in die neue Relation übernommen. Wie oben beschrieben werden einige Attribute im Sinne der Aufgabenstellung weggelassen. Die Datentypen einiger Attribute ändern sich im Zuge der Überführung. Pseudonyme (bzw. Patientenlaufnummern) welche in den Aufenthaltsdaten vorkommen, jedoch nicht in den Stammdaten werden aus der Kohorte mittels entsprechenden Vermerkes in der Relation work_kohorte_ausschluss ausgeschlossen.

Zu den Datensätzen wird weiters ein neues Attribut hinzugefügt (aufenthalt_ketten_nr) welches zusammenhängende Spitalsaufenthalte beschreibt. Die Daten aus der Relation import.aufenth_kette werden verwendet um dieses Attribut mit Werten zu füllen. Es wird jedem Aufenthalt dieselbe Nummer für das Attribut aufenthalt_ketten_nr vergeben sofern diese Beziehung (definiert als zusammenhängende Aufenthalte in der Relation aufenth_kette) in den Rohdaten beschrieben wird (über die Begin- und Enddaten der Kette). Es wird geprüft ob Überschneidungen dieser Aufenthaltsketten vorliegen, dies ist jedoch in den vorhandenen Daten nicht der Fall. Jedem Aufenthalt wird eine aufenthalt_ketten_nr vergeben (auch wenn diese nur von einem einzigen Aufenthalt verwendet wird).

In einem weiteren Schritt werden die inländischen Postleizahlen auf die aktuell verfügbaren Postleitzahlen in Österreich konvertiert. In den Rohdaten befinden sich beispielsweise Postleitzahlen zu Postfächern und veraltete Postleitzahlen. Diese werden im Script 142_lei_aufenthalt_cleanPlz.sql angepasst.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 7 veranschaulicht.

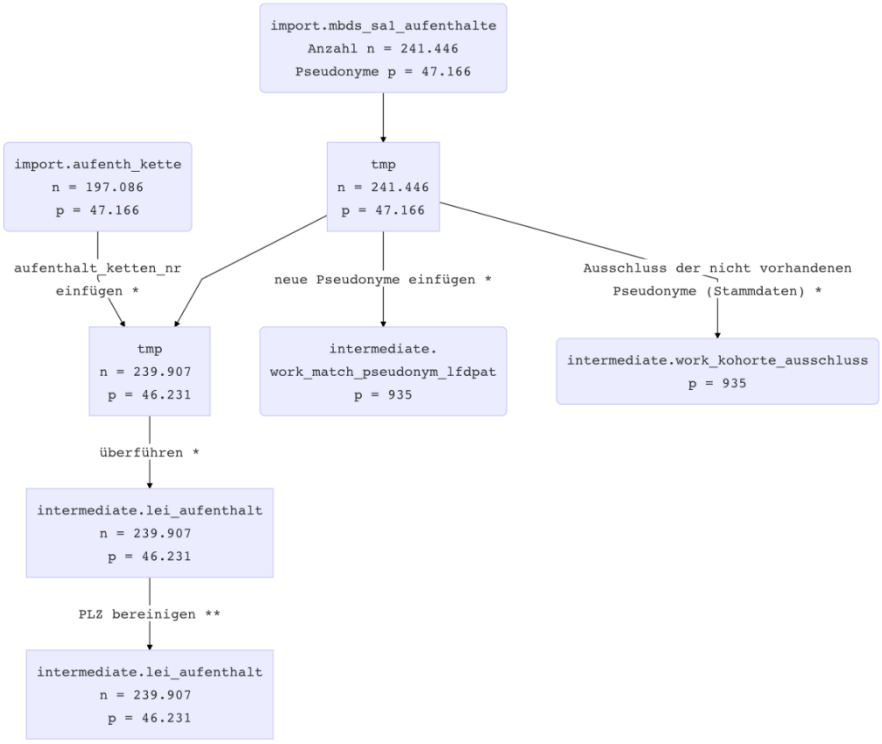

Abbildung 7: Überführung der Daten zu Aufenthalten
Skripte 141_lei_aufenthalt_transfer.sql (*) und 142_lei_aufenthalt_cleanPlz.sql (**)

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_aufenthalt
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

jahr	*ka_nr*
2015	K720
2014	K427
2015	K714
2012	K382
2014	K303

Table continues below

aufenthalt_nr	lfd_pat
0518d59cd130d2c92f3491927acaae7bee17a59344e9df6afa790dcd5d110b62	3125
0314073128	1085

b54b27631fc28cae44c0c851bc53e7d260cd5b87fc8ac94be6a0261880dde000	34576
2012246084	16625
1420000012	36652

Table continues below

aufenthalt_von	aufenthalt_bis	aufenthalt_belagstage
2015-07-30	2015-08-06	7
2014-06-11	2014-06-20	9
2015-07-23	2015-08-03	11
2012-11-05	2012-11-26	21
2014-01-01	2014-01-01	0

Table continues below

aufenthalt_art_aufnahme	aufenthalt_art_entlassung	pat_alter_entlassung
A	T	75
A	E	81
A	E	70
A	E	85
A	T	73

Table continues below

pat_geschlecht	pat_staatsbuerger	pat_plz	pat_plz_ausland	pat_land
W	AT	6130	-	AT
W	AT	4770	-	AT
M	AT	9990	-	AT
W	AT	3160	-	AT
M	AT	3321	-	AT

lkf_kostentraeger	lkf_ldf_gruppe	lkf_ldf_knoten	aufenthalt_ketten_nr
18	HDG01.32	A	151979
50	MEL26.02	A	135395
18	MEL21.01	C	96672
12	MEL10.01	A	156939
05	HDG06.03	D	165368

```
/* distinct_patient_count vs. row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_aufenthalt;
```

distinct_patient_count	row_count
46231	239907

```
/* aufenthalt_hauptdiagnose_count vs. aufenthalt_count
     (there are "aufenthalt" without "hauptdiagnose")  */
SELECT (
    SELECT COUNT(*) AS aufenthalt_mit_hauptdiagnose
        FROM intermediate.lei_aufenthalt_diagnose
```

```
        WHERE diagnose_typ LIKE 'H'
), (
    SELECT COUNT(*) AS aufenthalt_gesamt
    FROM intermediate.lei_aufenthalt_diagnose
);
```

aufenthalt_mit_hauptdiagnose	aufenthalt_gesamt
239805	864672

```
/* no. lei_aufenthalt / year */
SELECT jahr, COUNT(*) as lei_aufenthalt_count
    FROM intermediate.lei_aufenthalt
    GROUP BY jahr;
```

jahr	lei_aufenthalt_count
2011	21607
2012	39096
2013	43087
2014	44013
2015	44014
2016	48090

```
/* most common aufenthalt_art_aufnahme */
SELECT aufenthalt_art_aufnahme AS aufnahmeart, COUNT(aufenthalt_art_aufnah
me)
    FROM intermediate.lei_aufenthalt
    GROUP BY aufenthalt_art_aufnahme
    ORDER BY COUNT(aufenthalt_art_aufnahme) DESC;
```

aufnahmeart	count
A	188184
T	27928
K	19942
W	3104
H	743
L	4
R	2

```
/* average number of days in hospital */
SELECT ROUND(AVG(aufenthalt_belagstage), 2) AS avg_aufenthalt_belagstage
    FROM intermediate.lei_aufenthalt;
```

avg_aufenthalt_belagstage
5.92

3.5.5 Tabelle `lei_aufenthalt_diagnose`

Einem Spitalsaufenthalt zugeordnete Haupt- und Nebendiagnosen welche lt. ICD-10 BMG 2017 [4] codiert sind. Gesamt sind 864.672 Diagnosen zu 239.826 verschiedenen Aufenthalten enthalten.

Tabelle:	lei_aufenthalt_diagnose
Beschreibung:	Beinhaltet Datensätze zu Diagnosen (zugehörig zu einem Spitalsaufenthalt)
Anzahl Einträge:	864.672
PK:	jahr, ka_nr, aufenthalt_nr, diagnose_icd
FK:	jahr, ka_nr, aufenthalt_nr -> lei_aufenthalt, lfd_pat -> lei_stammdaten, diagnose_icd -> mat_icd10bmg, diagnose_vtr -> mat_kostentraeger
Anmerkung:	entspricht Tabelle mbds_sa2_diagnose in Rohdaten
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/150_lei_aufenthalt_diagnose_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/151_lei_aufenthalt_diagnose_transfer.sql

Diese Relation enthält die zu einem Spitalsaufenthalt vermerkten Diagnosen. Es ist zwischen Haupt- und Nebendiagnosen im Sinne der LKF-Abrechnung zu unterscheiden. Einem Aufenthalt ist immer genau eine Hauptdiagnose und beliebig viele Nebendiagnosen zugeordnet. In der Relation erfolgt die Zuordnung einer Diagnose zu einem Aufenthalt über die entsprechende Beziehung über die Attribute jahr, ka_nr, aufenthalt_nr.

Attribut	Typ	Verweis	Bemerkung
jahr	Integer	PK, FK	Jahr des Aufenthaltes
ka_nr	Varchar(4)	PK, FK	Kennung des jeweiligen Krankenhauses
aufenthalt_nr	Varchar(100)	PK, FK	Innerhalb eines Jahres und Krankenhauses eindeutige Nummer eines Aufenthaltes
lfd_pat	Integer	NN, FK	Eindeutige Kennung des Individuums
diagnose_typ	Varchar(1)	NN	Haupt- oder Zusatzdiagnose (H, Z)
diagnose_icd	Varchar(6)	FK	Diagnose codiert lt. ICD-10 BMG 2017
diagnose_vtr	Varchar(2)	FK	Versicherungs-/Kostenträger

Überführen der Rohdaten:

Die Daten werden ähnlich den ursprünglichen Daten in die neue Relation übernommen. Wie oben angemerkt werden einige Datentypen angepasst und Beziehungen zu anderen Relationen eingeführt.

Es existieren Diagnosen zu Aufenthalten welche nicht durch die Relation lei_aufenthalte beschrieben werden. Diese Diagnosen werden aus der Verarbeitung ausgeschlossen, da kein Kontext zu einer Behandlung, bzw. zu einem Krankenhausaufenthalt herstellbar ist. Bei sieben der 239.826 Aufenthalte (von 46.231 Individuen) existiert nur eine Nebendiagnose ohne zugeordneter Hauptdiagnose. Auf eine gesonderte Behandlung dieser wird aufgrund der geringen Anzahl verzichtet. Da die

Einschlusskriterien für die spätere Betrachtung der Qualitätsindikatoren stets über die Hauptdiagnose erfolgen ist ein Ausschluss auch nicht erforderlich.

Die initiale Vermutung dass sich in den Rohdaten Diagnosecodes des ICD-9 Systems befinden, bewahrheitet sich nach genauerer Analyse nicht. Die rein nummerischen Codes stellen österreichspezifische Codes lt. BMG dar. Um eine einheitliche Datenlage zu erhalten werden alle enthaltenen Diagnosen auf die ICD-10 Systematik (herausgegeben vom Bundesministerium für Gesundheit und Frauen in der Version BMGF-Version 2017) konvertiert. Die verwendeten Codes sind seitens des Ministeriums definiert und sind als 5 oder 6-stelliger Code angegeben.

Die zuvor als vermeintlichen ICD-9 Codes aufgetretenen Werte stellen in Wirklichkeit ICD-10 BMG 2017 spezifische Codes dar. In der zugehörigen Dokumentation werden unter Kapitel XXa nummerische Codes definiert welche syntaktisch denen des ICD-9 ähnlich sind. Diese Codes beschreiben Revisionsgründe in der Endoprothetik bzw. exogene Noxen der Ätiologie. Da sich die vorgefundenen Daten genau auf diese Codes beschränken und zudem nur als Nebendiagnosen deklarierte Diagnosen diese Ausprägung aufweisen werden alle Daten in `lei_aufenthalt_diagnose` als ICD-10 BMG 2017 [4] Diagnosen betrachtet. Zum Zwecke der Konsistenzprüfung wird eine Fremdschlüsselbeziehung des Attributs `diagnose_icd` auf die Tabelle `mat_icd10bmg` eingeführt.

Einige der vorkommenden Diagnosecodes stellen keine gültigen Codes im Sinne des ICD-10 BMG 2017 dar, da diese beispielsweise ein Überkapitel, also die dreistellige allgemeine Systematik, benennen. Dies trifft auf 61 der 6.638 Diagnosen und 34.329 der 864.724 Datensätze der Relation zu. Diese Datensätze werden manuell angepasst. Durch diese Anpassungen verletzten einige Datensätze die Schlüsselbedingung und würde mehrfach einem Aufenthalt zugeordnet werden. Diese Datensätze (welche ohnehin keine zusätzliche Informationsgewinn darstellen) werden daher aus der Verarbeitung ausgeschlossen und gelöscht. Die Zuordnung der dreistelligen Systematik auf eine differenziertere Codierung erfolgt in einem mehrstufigen Verfahren und beginnt mit der Einordnung in vorhandene Unterkategorien mit der Endung *'.9 – nicht näher bezeichnet'* falls diese im jeweiligen Kapitel vorhanden sind. Einer manuellen Bearbeitung unterliegen 61 der 6.638 Diagnosen im Datensatz. Die nach der Einteilung in die oben genannte Kategorie verbleibenden Codes werden manuell unter Beachtung der dokumentierten Ein-/Ausschlusskriterien in Unterkategorien eingeordnet. Diese Kriterien sind in der Dokumentation des ICD-10 BMG 2017 definiert. So wird beispielsweise der in der aktuellen Version nicht mehr vorhandene Code 'I84 Hämorrhoiden' dem vorhandenen Code 'K64.9 Hämorrhoiden, nicht näher bezeichnet' zugeordnet.

Die Einteilung der betreffenden Codes kann in der zugehörigen Datei (siehe `DEXHELPP-GitLab: data/lei_aufenthalt_diagnose_ICD10_BMG_mapping.xlsx`), inklusive einer Begründung und einem Verweis auf die verwendeten Ressourcen, nachvollzogen werden. Die Änderungen werden mit dem SQL-Skript (siehe `DEXHELPP-GitLab: schema_modify/152_lei_aufenthalt_diagnose_cleanIcd.sql`) auf die Datenbank angewendet.

Der Vollständigkeit halber sei erwähnt: Zusätzlich zu den oben genannten Anpassungen werden zwei Diagnosen unter Beachtung der Patientenhistorie manuell zugeordnet. Hierbei handelt es sich um Codierungen zu 'Vorzeitige Wehen und Entbindung' mit der Angabe ob eine Entbindung stattgefunden hat oder nicht. Diese Änderung betrifft

lediglich zwei Einträge und wird zum Zwecke der Schaffung einer einheitlichen Datenbasis durchgeführt.

Nachfolgend wird der gesamte Prozess der Datenüberführung in Abbildung 8 veranschaulicht.

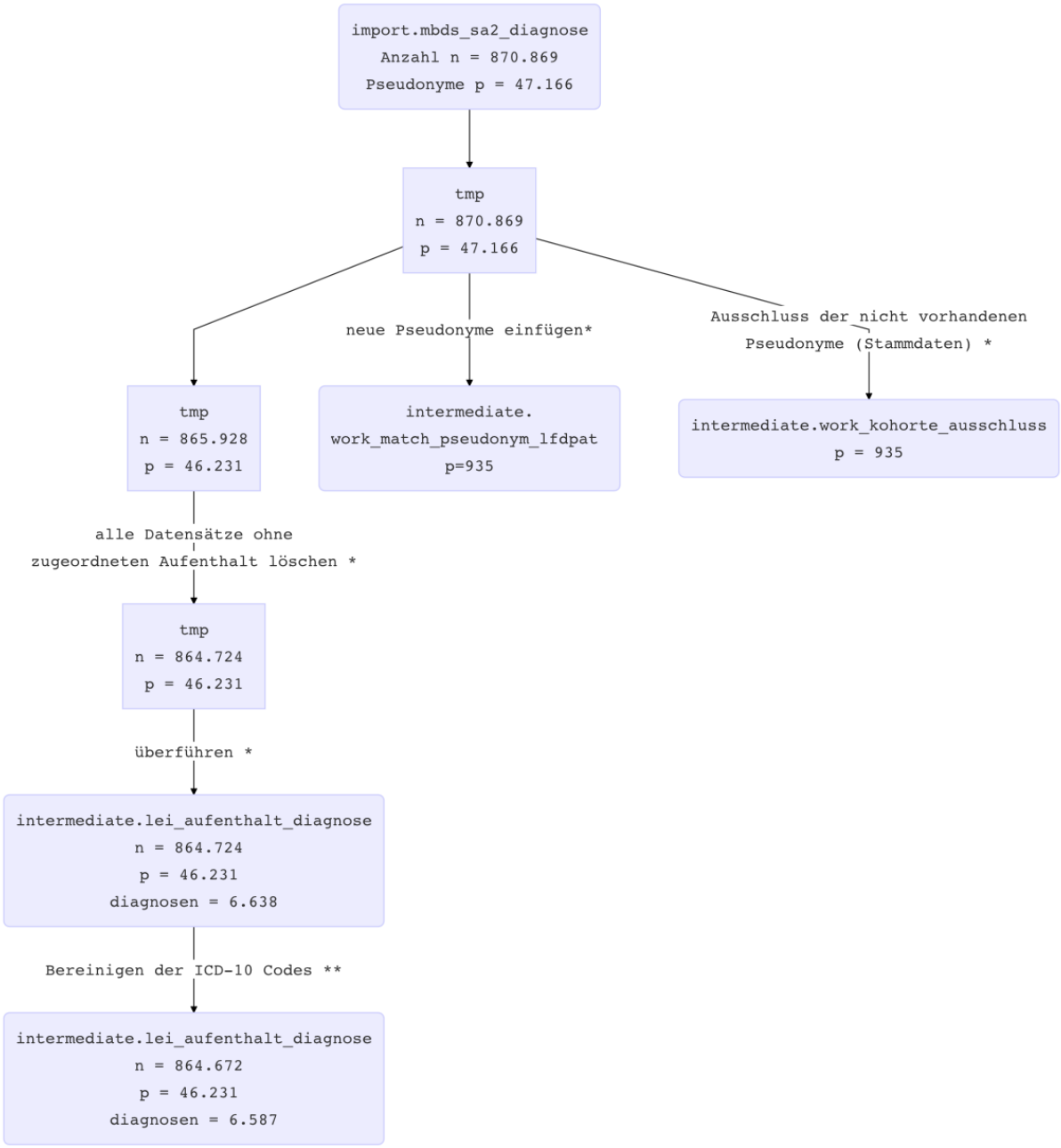

*Abbildung 8: Überführung der Daten zu den einem Aufenthalt zugeordneten Diagnose Skripte 151_lei_aufenthalt_diagnose_transfer.sql (*) und 152_lei_aufenthalt_diagnose_cleanIcd.sql (**) bzw. 153_lei_aufenthalt_diagnose_fk.sql (*** nicht gesondert aufgeführt)*

```
/* 5 random entries */
SELECT *
```

```
    FROM intermediate.lei_aufenthalt_diagnose
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

jahr	*ka_nr*
2015	K912
2014	K956
2015	K431
2016	K608
2016	K704

Table continues below

aufenthalt_nr	lfd_pat
79101358db36b423c3134e88f11c8b0428343b3fbe93fb89dbf39bdbb36d13ce	14781
5614121417	40273
384a1e31675d9da4e37de38b3ce9a9733d9bbf079f0dfba66fd6710103069e27	35297
664d30e90560b5580962a19dbeb99d774a3ae7fe09362640722e0de15cbb70d9	29139
4bc84aeb375a74e784c83de5651f1be4306f33b4b6ba28b73916c60e5ed3b8b4	11975

diagnose_typ	*diagnose_icd*	diagnose_vtr
Z	I10	11
Z	F32.9	-
Z	I63.4	15
Z	I67.2	15
H	C18.24	18

```
/* distinct_patient_count vs. Row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_aufenthalt_diagnose;
```

distinct_patient_count	row_count
46231	864672

```
/* main vs. Secondary diagnoses count */
SELECT diagnose_typ, COUNT(diagnose_typ)
    FROM intermediate.lei_aufenthalt_diagnose
    GROUP BY diagnose_typ
    ORDER BY COUNT(diagnose_typ) DESC
    LIMIT 10;
```

diagnose_typ	count
Z	624867
H	239805

```
/* most common diagnoses */
SELECT diagnose_icd, COUNT(diagnose_icd)
    FROM intermediate.lei_aufenthalt_diagnose
    GROUP BY diagnose_icd
    ORDER BY COUNT(diagnose_icd) DESC
    LIMIT 10;
```

diagnose_icd	count
I10	88257
I25.1	39893
I21.4	31243
I25.9	29409
E11.9	20741
E78.0	16990
I21.9	16823
E78.5	15405
I48.9	13451
N18.9	10362

3.5.6 Tabelle `lei_aufenthalt_mel`

Im Rahmen eines Spitalaufenthalts erbrachte medizinische Einzelleistungen codiert lt. Leistungskatalog 2018 des BMGF werden in dieser Relation abgebildet. Es sind gesamt 498.843 Einträge zu 169.242 verschiedenen zugeordneten Aufenthalten vorhanden

Tabelle:	`lei_aufenthalt_mel`
Beschreibung:	Beinhaltet Datensätze medizinischen Einzelleistungen (zugehörig zu einem Spitalsaufenthalt)
Anzahl Einträge:	498.843
PK:	`lfd_mel`
FK:	`jahr, ka_nr, aufenthalt_nr` -> `lei_aufenthalt`, `lfd_pat` -> `lei_stammdaten`, `ka_nr` -> `mat_krankenanstalten`, `mel_code` -> `mat_mel_kal`, `mel_vtr` -> `mat_kostentraeger`
Anmerkung:	entspricht Tabelle `mbds_sa3_mel` in Rohdaten
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/160_lei_aufenthalt_mel_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/161_lei_aufenthalt_mel_transfer.sql`

Diese Relation enthält medizinische Einzelleistungen welche einem Krankenhausaufenthalt zugeordnet sind. Pro Aufenthalt können beliebig viele MELs anfallen. Daher wird zur Identifizierung innerhalb der Relation eine `lfd_mel` eingeführt. Die Datensätze lassen sich weiterhin über die Attribute `jahr`, `ka_nr`, `aufenthalt_nr` einem Aufenthalt zuordnen.

Attribut	Typ	Verweis	Bemerkung

lfd_mel	Serial	PK	Eindeutige Kennung des Eintrages
jahr	Integer	FK, NN	Jahr des Aufenthaltes
ka_nr	Varchar(4)	FK, NN	Kennung des jeweiligen Krankenhauses
aufenthalt_nr	Varchar(100)	FK, NN	Innerhalb eines Jahres und Krankenhauses eindeutige Nummer eines Aufenthaltes
lfd_pat	Integer	FK, NN	Eindeutige Kennung des Individuums
mel_code	Varchar(5)	FK, NN	Code lt. Leistungskatalog Sozialministerium (Code für ausgewählte medizinische Einzelleistungen), siehe Beschreibung der Tabelle mat_mel_kal
mel_pat_seite	Varchar(1)	-	Körperseite (L,R,NULL)
mel_anzahl_leistungen	Integer	NN	Anzahl der Einzelleistungen
mel_datum	Date	NN	Datum der Leistungserbringung
mel_vtr	Varchar(2)	FK	Versicherungsträger/Kostenträger

Überführen der Rohdaten:

Die Daten werden ähnlich den ursprünglichen Daten in die neue Relation übernommen. Es wird ein Attribut (lfd_mel) eingeführt um die Datensätze eindeutig identifizierbar zu machen. Im Rohdatensatz sind für die Richtungsbezeichnung der Mel Codes neben den Werten 'L', 'R' und NULL auch der Wert '-' vorhanden; dieser wird entfernt und durch NULL ersetzt.

In den Stammdaten nicht vorkommende Individuen werden, wie in den vorhergehenden Absätzen beschrieben, ausgeschlossen. Ebenso wird mit Einträgen ohne zugehörigen Aufenthalt umgegangen.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 9 veranschaulicht.

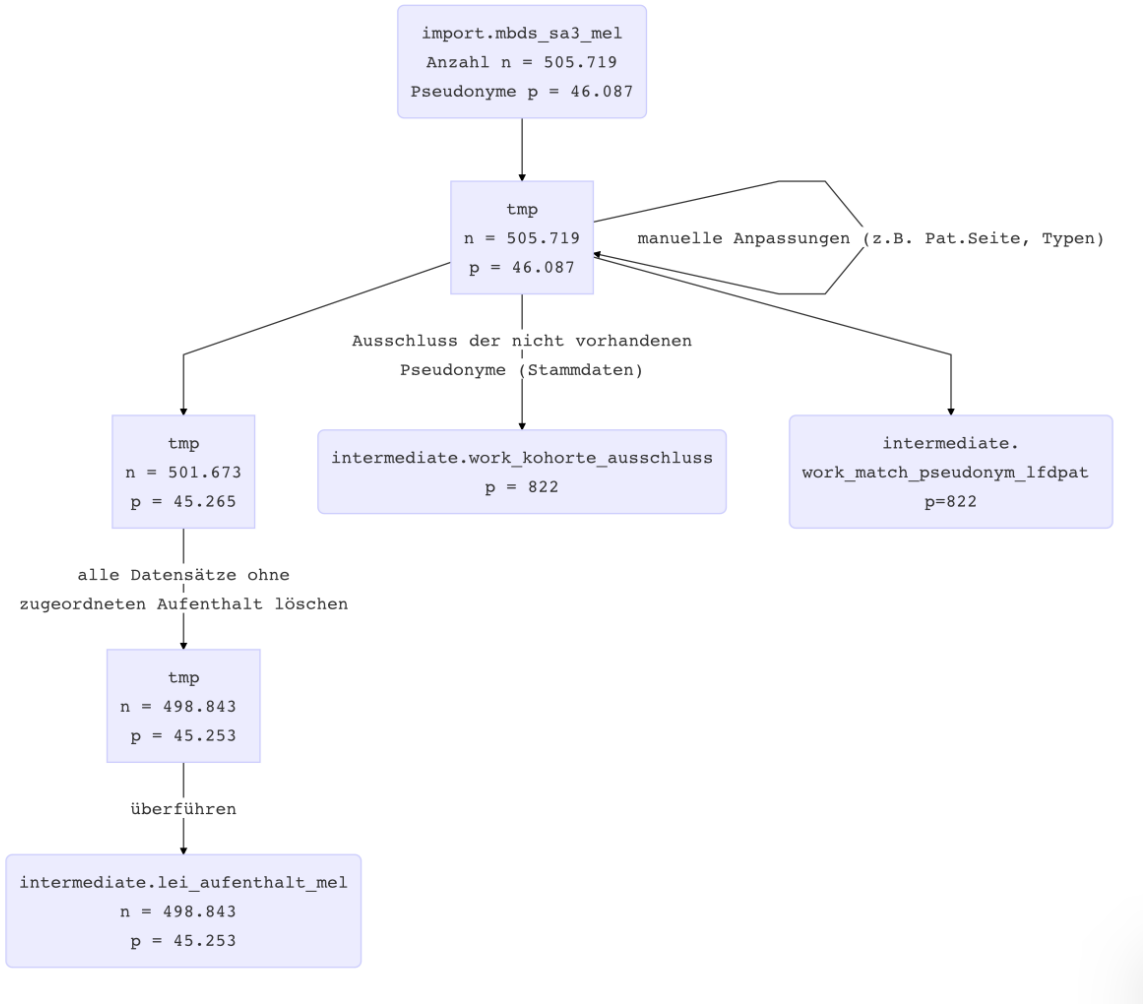

Abbildung 9: Überführung der einem Aufenthalt zugeordneten MEL-Codes

Skript 161_lei_aufenthalt_mel_transfer.sql

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_aufenthalt_mel
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

lfd_mel	jahr	ka_nr
296583	2012	K722
79184	2013	K612
287863	2012	K528
75687	2013	K534
4329	2015	K416

Table continues below

aufenthalt_nr	lfd_pat
2012037953	30010

		2013420104		444
		2265095512		15674
		0013047123		30545
		8be62cd5ca9ce6aac1bd9cfeae902cff2df5a50ae76ea13387481a2bf7955d1e		9998
mel_code	mel_pat_seite	mel_anzahl_leistungen	mel_datum	mel_vtr
ZB010	-	1	2012-07-17	18
DD060	-	1	2013-10-03	15
ZA010	-	1	2012-03-12	16
DD010	-	1	2013-05-14	17
ED054	R	1	2015-05-07	14

```
/* distinct_patient_count vs. Row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_aufenthalt_mel;
```

distinct_patient_count	row_count
45253	498843

```
/* most common mel_codes*/
SELECT mel_code, COUNT(mel_code)
    FROM intermediate.lei_aufenthalt_mel
    GROUP BY mel_code
    ORDER BY COUNT(mel_code) DESC
    LIMIT 10;
```

mel_code	count
DD010	54690
PE010	42695
DD040	38254
DD060	37583
ZA010	25034
ZC010	14138
FV045	13398
ZB010	11010
BF020	8548
ZN360	7292

```
/* average mel_anzahl_leistungen per entry */
SELECT ROUND(AVG(mel_anzahl_leistungen), 2) AS avg_number_of_mel
    FROM intermediate.lei_aufenthalt_mel;
```

avg_number_of_mel
1.11

3.5.7 Tabelle `lei_leistung`

Leistungen des niedergelassenen Bereiches welche für ein Individuum über den jeweiligen Versicherungsträger abgerechnet wurde. Es sind 18.046.490 Leistungen von

46.039 verschiedenen Patienten enthalten. Die Leistungen der einzelnen Versicherungsträger werden mittels eines Metamodells abstrahiert (siehe Tabelle mat_meta_leistung) und medizinischen Einzelleistungen und KAL-Codes zugeordnet (siehe Tabelle mat_meta_leistung_mel_kal und Tabelle mat_mel_kal).

Tabelle:	lei_leistung
Beschreibung:	Beinhaltet Leistungen aus dem extramuralen Bereich
Anzahl Einträge:	18.046.490
PK:	lfd_leistung
FK:	lfd_pat -> lei_stammdaten, leistung_vtr -> mat_kostentraeger, leistung_fachgruppe -> mat_fachgruppe, leistung_meta_code -> mat_meta_leistung
Anmerkung:	entspricht Tabelle ek in Rohdaten
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/180_lei_leistung_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/181_lei_leistung_transfer.sql

Diese Relation enthält Leistungen aus dem extramuralen Bereich. Die erbrachten Leistungen werden von unterschiedlichen Versicherungsträgern unterschiedlich kodiert. Eine Vereinheitlichung mittels Daten aus der GAP-DRG2 Datenbank wurde angedacht scheiterte aber an den verfügbaren Daten in dieser. Eine Zuordnung der einzelnen Leistungen zu entsprechenden MEL bzw. KAL Codes wird letztlich mit einer Datenbank der NÖGKK (Ansprechperson: Hr. Rohbausch) erreicht, welche Informationen über die Leistungspositionen der einzelnen Versicherungsträger enthält. Aus dieser Datenbank werden im weiteren Verlauf auch abstraktere Meta-Leistungspositionen generiert (siehe Tabelle mat_meta_leistung). Die in den Rohdaten enthaltenen Leistungspositionen werden in weiterer Folge auf diese Metaleistungen verweisen. Das Attribut leistung_meta_code enthält diese Zuordnung zu Metaleistungen und dadurch zu den zugehörigen MEL bzw. KAL Codes (siehe unten).

Attribut	Typ	Verweis	Bemerkung
lfd_leistung	Serial	PK	Eindeutige Identifizierung des Datensatzes
lfd_pat	Integer	NN, FK	Eindeutige Kennung des Individuums
leistung_datum	date	NN	Tag der Leistungserbringung
leistung_nr	varchar(50)	NN	Leistungsnummer, pro Versicherungsträger (vgl. Attribut posnr in den Rohdaten)
leistung_bez	Text	-	Textuelle Beschreibung der Leistung
leistung_honoid	Integer	-	honoId beschreibt die Leistung, individuell per Versicherungsräger und auch hier teilweise nicht eindeutig. Dieses Attribut wird verwendet um die Metaleistungen zuzuordnen.

leistung_anzahl	Numeric	-	Anzahl der bezogenen Leistungen
leistung_betrag	Numeric	-	zugeordnete monetärer Betrag der Leistung
leistung_vtr	varchar(2)	NN, FK	Versicherungsträger
leistung_fachgruppe	Integer	FK	Fachgruppe des Erbringers
leistung_kategorie	Integer		unbek. Verwendung
leistung_meta_code	Integer	FK	Zuordnung zu einer Metaleistung aus der Relation mat_meta_leistung und dadurch indirekt zu MEL bzw. KAL Codes

Das Attribut leistung_fachgruppe enthält codierte Werte zu den Fachgruppen der Leistungserbringer (beispielsweise entspricht Fachgruppe 1 einem Arzt für Allgemeinmedizin). Die Codes wurden durch Daten aus verschiedenen Quellen erklärt (GAP-DRG Wiki, Projekt ADE-PIM). Die Codes zu vorhandenen Fachgruppen werden im GitLab Repository unter *data/mat_fachgruppe.xlsx* gesammelt beschrieben.

Meta-Leistungen

Der Problematik verschiedener Leistungscodierungen und -bezeichnungen der einzelnen Versicherungsträger wird mit einer Zuordnung der Leistungspositionen zu einem einheitlichen Codierschema begegnet. Diese Maßnahme ist erforderlich um Auswertungen zu den einzelnen medizinisch relevanten Leistungen zu ermöglichen. Diese Zuordnung erfolgt letztlich auf Basis eines Datensatzes der NÖGKK (Zugesendet per Mail durch Hrn. Rohbausch am 02.08.2018, siehe data/mat_meta_leistung/src/*). Dieser Datensatz enthält Zuordnungen der Leistungen der einzelnen Versicherungsträger auf ein gemeinsames Metaleistungsmodell. Die Datenqualität dieses Datensatzes ist nicht optimal da, lt. Rücksprache mit der NÖGKK, zum Teil Meldungen der Versicherungsträger fehlen. Dieser Datensatz ist weiters die Grundlage für die Synthese der Relationen mat_meta_leistung und mat_meta_leistung_mel_kal (auf welche von der vorliegenden Relation auch verwiesen wird, vgl. Tabelle mat_meta_leistung, Tabelle mat_meta_leistung_mel_kal).

Die Zuordnung unserer Leistungspositionen zu diesen Metaleistungen erfolgt in einem mehrstufigen Verfahren und wird durch die oben genannten Datensätze unterstützt. Die Anwendung auf die Daten kann in der Datei schema_modify/187_lei_leistung_zuordnung_meta.sql nachvollzogen werden. Folgende Schritte werden zur Zuordnung durchgeführt:

1. Zuordnung von Leistungen anhand des teilweise vorhandenen Attributs leistung_honoid. Dies stellt eine exakte Zuordnung dar und betrifft die ersten 14.126 von 44.320 Positionen
2. Zuordnen von Leistungen anhand der exakten Übereinstimmung des Attributs leistung_bez der Daten und TP_Leistungstext des Vergleichsdatensatzes. Dies stellt ebenfalls eine exakte Zuordnung dar und betrifft weitere 14.647 von 44.320 Positionen.
3. Zuordnen von Leistungen desselben Versicherungsträgers (leistung_vtr der Daten entspricht vstr der Vergleichsdaten) anhand der leistung_nr der Daten und dem fachschlüssel der Vergleichsdaten. Dies sollte ebenfalls eine exakte Zuordnung darstellen und betrifft weitere 2.100 der 44.320 Positionen.

4. Zuordnen von Leistungen anhand der Übereinstimmung von `leistung_bez` und `TP_Leistungstext` wie in Schritt 2 beschrieben. Da jedoch einige verschiedene Schreibweisen und Textcodierungsfehler im Datensatz vorliegen, sowie teilweise Texte abgeschnitten sind wird mit der zusätzlichen Einschränkung auf die ersten 40 Zeichen und das Entfernen von Leerzeichen, Umlauten und weiteren Sonderzeichen gearbeitet. Dadurch können weitere 4.112 von 44.320 Positionen zugeordnet werden.
5. Nach den oben genannten Schritten verbleiben 10.006 Positionen, welche nicht zugeordnet werden konnten und in weiterer Folge mit Fuzzy String Matching Methoden bearbeitet werden. Dazu werden die verbleibenden zu bearbeitenden Positionstexte (`leistung_bez` aus den Daten, ohne Sonderzeichen, etc.) aus der Datenbank exportiert und mit dem Python Package difflib und der Funktion SequenceMatcher der zuordenbaren Menge zugeordnet. Die Zuordnung ergibt ein Matching zwischen `leistung_bez` und `TP Leistungstext` und einem zugehörigen Ähnlichkeitsmaß. Intern wird der 'gestalt pattern matching'-Algorithmus von Ratcliff und Obershelp verwendet. Die erste Zuordnung erfolgt für alle Versicherungsträger getrennt und nach visueller Inspektion wird ein Threshold eingeführt ab welchem die Ergebnisse verwertet werden. Die Ergebnisse und Rohdaten sowie das zugehörige Skript sind im Projektverzeichnis unter `sql_playground/2018-08-03_metahono_infos_von_rohbausch/v3/*` zu finden. Auf diese Weise werden weitere 3.778 der 44.320 Positionen zugeordnet.
6. Wie in Punkt 5 beschrieben wird dasselbe Verfahren nochmals angewendet, jedoch ohne die Einschränkung eine Zuordnung innerhalb eines Versicherungsträgers zu erreichen. Die Ressourcen dazu sind unter `sql_playground/2018-08-03_metahono_infos_von_rohbausch/v4/*` zu finden. Auf diesem Weg werden weitere 2.724 der 44.320 Positionen zugeordnet.
7. Es bleiben schließlich 2.833 Positionen ohne Zuordnung übrig.

Da einige der oben genannten Berechnungen rechenaufwändig sind und einige Zeit in Anspruch nehmen (insbesonders die Ausführung der Python Skripts) werden die Zuordnungen statisch in `schema_modify_185_lei_leistung_zuordnung_meta_tmpv3.sql` und `schema_modify_186_lei_leistung_zuordnung_meta_tmpv4.sql` abgelegt und auf die Daten später angewendet. Alle restlichen oben beschrieben Berechnungen finden sich in `schema_modify/187_lei_leistung_zuordnung_meta.sql`.

Überführen der Rohdaten:

Die Daten werden ähnlich den ursprünglichen Daten in die neue Relation übernommen. Es wird ein Feld `lfd_leistung` neu eingeführt um die Datensätze eindeutig identifizierbar zu machen. Datensätze zu Pseudonymen welche nicht in den Stammdaten vorkommen werden entfernt. Zusätzlich werden die oben angeführten Zuordnungen zu Metaleistungen eingefügt.

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 10 veranschaulicht.

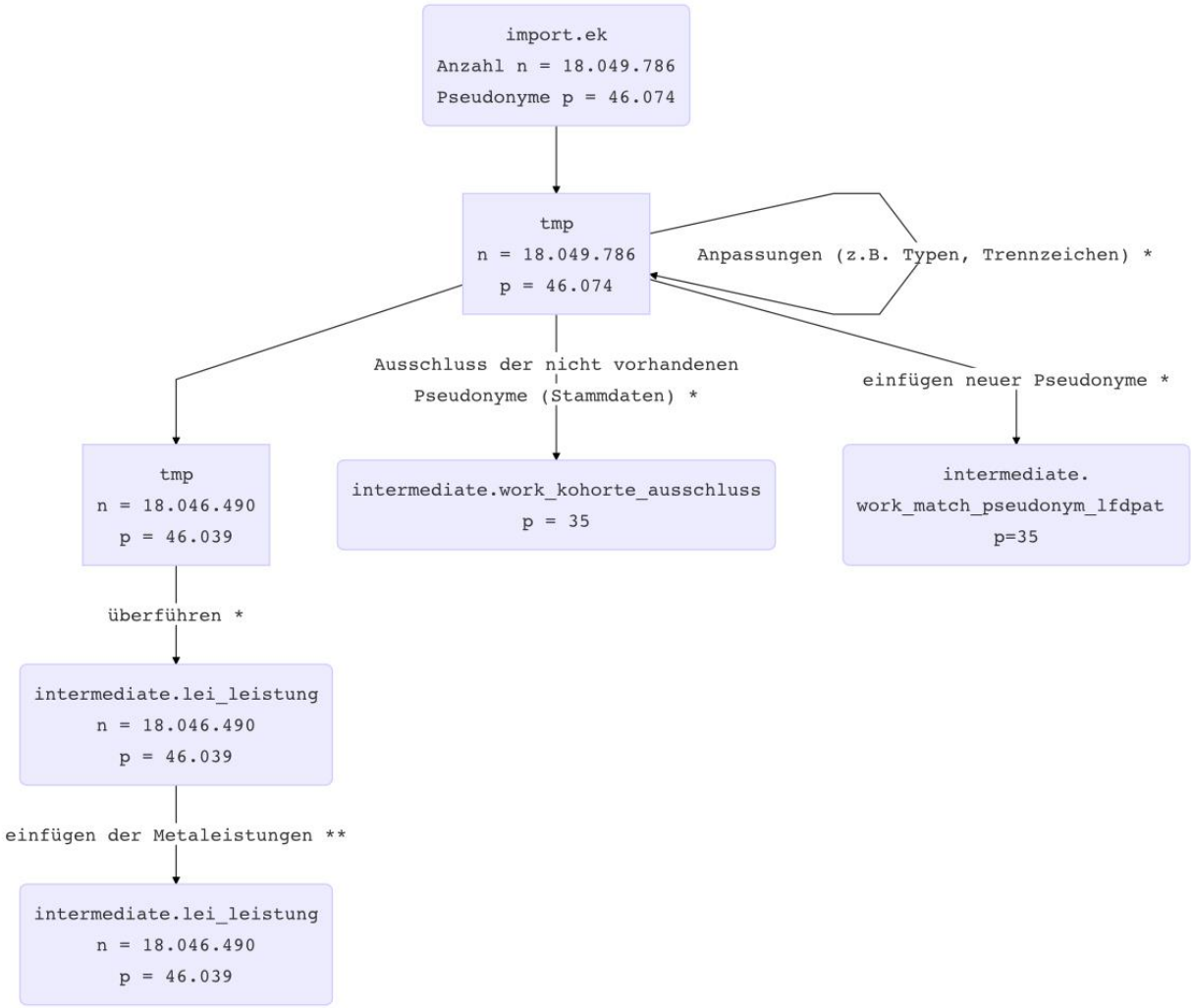

Abbildung 10: Überführung der Daten zu Leistungen
Skripte 181_lei_leistung_transfer.sql (*) und 187_lei_leistung_zuordnung_meta.sql (**)

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_leistung
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

lfd_leistung	lfd_pat	leistung_datum	leistung_nr
6889474	44972	2012-05-08	ZAM/INDI/50
6222622	29844	2015-12-22	ZAM/INDI/15
4946355	11383	2014-02-27	ZAM/INDI/1A
12258486	27460	2013-04-19	ZAM/INDI/361
15538338	33519	2011-07-01	09k

Table continues below

leistung_bez	leistung_honoid	leistung_anzahl
(Folge)ordination	8403	1

Ordination	20	1
TAGESORDINATION	6405	1
Grundvergütung	4965	1
Harnröhrenstriktur, Bougierung	-	1

Table continues below

leistung_betrag	leistung_vtr	leistung_fachgruppe	leistung_kategorie
6.2	17	1	99999
7.54	15	1	7000
3.71	14	1	800
3.41	12	1	99999
3.49	16	16	3

leistung_meta_code
10201
10201
10201
10101
130309

```
/* distinct_patient_count vs. row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_leistung;
```

distinct_patient_count	row_count
46039	18046490

```
/* most common leistung_nr */
SELECT leistung_nr, leistung_bez, COUNT(leistung_nr)
    FROM intermediate.lei_leistung
    GROUP BY leistung_nr, leistung_bez
    ORDER BY COUNT(leistung_nr) DESC
    LIMIT 10;
```

leistung_nr	leistung_bez	count
ZGRUND/INDI/12	Ordination	438905
ZAM/INDI/15	Ordination	386582
ZSTANDART/INDI/1	Ordination	331362
ZGRUND/INDI/9	Ordination eingeschränkt	319241
ZAM/INDI/1A	TAGESORDINATION	241409
015	Ordination	187963
0102010	Erste Ordination	184475
ZAVA/INDI/FP	Fallpauschale	135764
ZAM/INDI/FP	Fallpauschale	135007
ZAM/INDI/361	Grundvergütung	129492

```
/* average leistung_betrag */
SELECT ROUND(AVG(leistung_betrag), 2) AS avg_leistung_betrag
    FROM intermediate.lei_leistung;
```

avg_leistung_betrag
10.95

3.5.8 Tabelle `lei_heilmittel`

Beinhaltet Daten über durch den Versicherungsträger abgerechneten und einem Patienten zugeordneten Heilmittel. Bei pharmazeutischen Produkten wird die entsprechende Pharmazentralnummer zur Bezeichnung des Medikaments angegeben welche auf die entsprechende Relation (siehe Tabelle `mat_pharmazie`) verweist. In der Relation sind in Summe 9.412.985 verrechnete Heilmittel von 46.015 Patienten vorhanden.

Tabelle:	`lei_heilmittel`
Beschreibung:	Beinhaltet Datensätze zu bezogenen Heilmitteln (Medikamente)
Anzahl Einträge:	9.412.985
PK:	`lfd_heilmittel`
FK:	`lfd_pat` -> `lei_stammdatenheilmittel_pharma_nr` -> `mat_pharmazie`
Anmerkung:	entspricht Tabelle `hm` in Rohdaten
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/170_lei_heilmittel_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/171_lei_heilmittel_transfer.sql`

Diese Relation enthält von Patienten bezogene Heilmittel, im Wesentlichen sind dies Medikamente und deren Parameter. Die Medikamente werden durch eine Pharmazentralnummer identifiziert und in der Relation `mat_pharmazie` genauer beschrieben, inklusive der Stoffgruppe in Form des ATC-Codes. Die Beziehung beider Relationen wird über die Pharmazentralnummer hergestellt.

Attribut	Typ	Verweis	Bemerkung
`lfd_heilmittel`	Serial	PK	Identifikator eines Datensatzes
`lfd_pat`	Integer	NN, FK	Eindeutige Kennung des Individuums
`heilmittel_datum`	Date	NN	Tag der Leistungserbringung
`heilmittel_pharma_nr`	Integer	FK	Pharmazentralnummer des Heilmittels
`heilmittel_anzahl`	Numeric	-	Anzahl der bezogenen Heilmittel
`heilmittel_betrag`	Numeric	-	Betrag der Heilmittel-Kosten

Überführen der Rohdaten:

Die Daten werden ähnlich den ursprünglichen Daten in die neue Relation übernommen. Es wird ein Feld `lfd_heilmittel` neu eingeführt um die Datensätze eindeutig identifizierbar zu machen. In den Rohdaten wurden unterschiedliche Dezimaltrennzeichen in `anzahl` und `betrag` gefunden, welche entsprechend korrigiert

werden. Daten zu Individuen welche nicht in den Stammdaten vorkommen werden entfernt.

Seitens der WHO wurden in den letzten Jahren einige ATC Codes abgeändert. Eine Zusammenfassung dieser Änderungen ist auf whocc.no[2] zu finden. Diese Änderungen beziehen sich auf die Codierung der Stoffgruppen und stellen einfache Umbenennungen dar, ohne dabei die Semantik des Codes zu verändern. Daher werden alle nicht aktuellen ATC-Codes auf die neueste Version lt. WHO abgeändert. Die ATC-Codes werden in weiterer Folge nicht in der vorliegenden Relation gespeichert, sondern zusammen mit Informationen über das jeweilige Heilmittel in der Relation `mat_pharmazie`. Fehlende Einträge in der Relation `mat_pharmazie` werden während der Überführung der Daten eingetragen (jedoch ohne weitere Attribute, dies entspricht also einem "Dummy"-Eintrag).

Nachfolgend wird der Prozess der Datenüberführung in Abbildung 11 veranschaulicht.

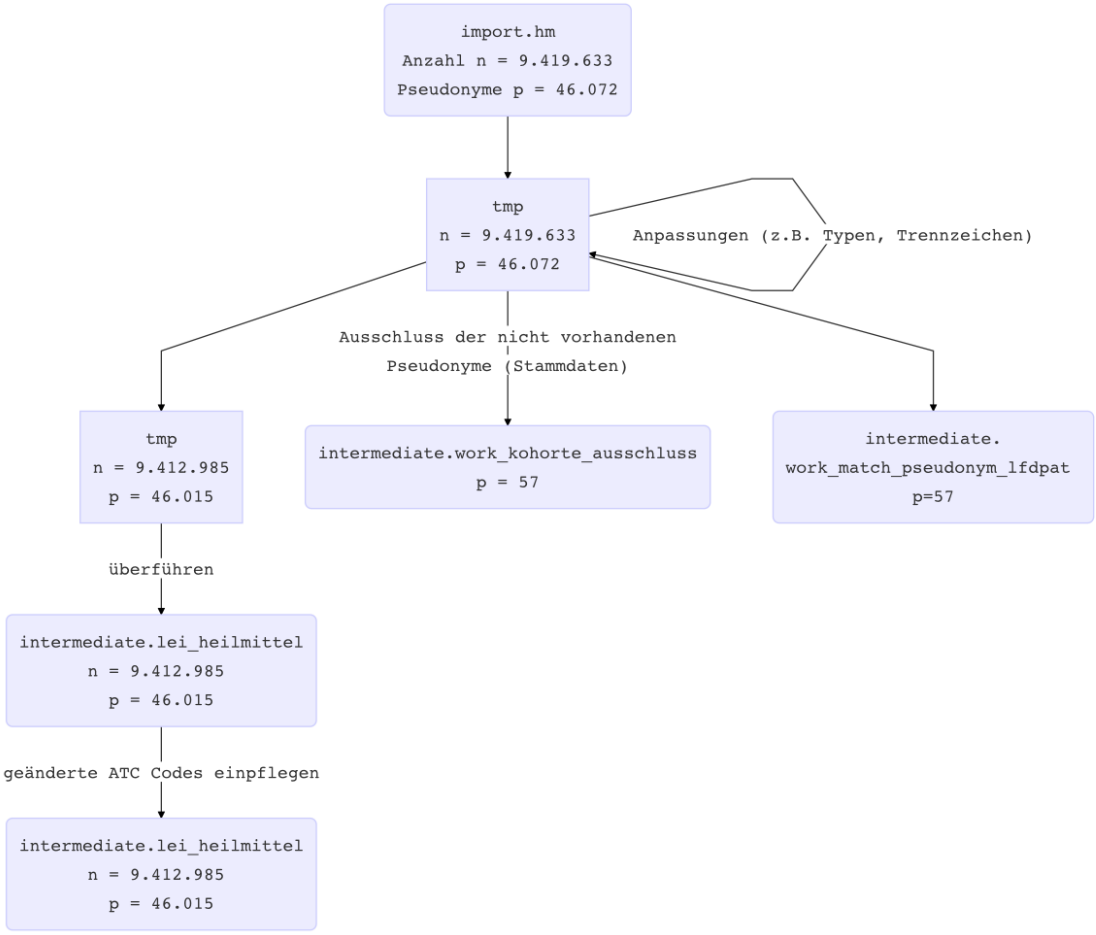

Abbildung 11: Überführung der Daten zu Heilmitteln
Skript 171_lei_heilmittel_transfer.sql

[2] https://www.whocc.no/atc_ddd_alterations__cumulative/atc_alterations/?order_by=2

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_heilmittel
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

lfd_heilmittel	lfd_pat	heilmittel_datum	heilmittel_pharma_nr
6141945	20082	2014-11-06	37960
8442430	13604	2016-03-16	1322125
6482383	44687	2016-02-02	2428245
7279745	936	2011-02-07	2443061
6237205	9115	2013-04-29	1130041

heilmittel_anzahl	heilmittel_betrag
1	8.85
1	3.3
2	12.9
1	17.05
1	8.9

```
/* distinct_patient_count vs. row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count, COUNT(*) AS row_
count
    FROM intermediate.lei_heilmittel;
```

distinct_patient_count	row_count
46015	9412985

```
/* most common heilmittel_pharma_nr */
SELECT heilmittel_pharma_nr, COUNT(heilmittel_pharma_nr)
    FROM intermediate.lei_heilmittel
    GROUP BY heilmittel_pharma_nr
    ORDER BY COUNT(heilmittel_pharma_nr) DESC
    LIMIT 10;
```

heilmittel_pharma_nr	count
1292648	208475
9999927	148764
1130041	142592
3771198	110106
2436227	103921
1326614	103442
1269566	93803
3538378	92395
2437528	73905
1339143	70112

```
/* average heilmittel_betrag */
SELECT ROUND(AVG(heilmittel_betrag), 2) AS avg_heilmittel_betrag
    FROM intermediate.lei_heilmittel;
```

avg_heilmittel_betrag
23.28

3.5.9 Tabelle `lei_arbeitsunfaehigkeit`

Beinhaltet Daten zu Arbeitsunfähigkeitsmeldungen von Versicherten an den jeweiligen Versicherungsträger samt eingetragener, als ICD-10 BMG 2017 [4]codierte Diagnose. Es sind 97.266 Datensätze von 12.894 Individuen vorhanden.

Tabelle:	`lei_arbeitsunfaehigkeit`
Beschreibung:	Beinhaltet die Beschreibungen von Krankenständen
Anzahl Einträge:	97.266
PK:	`lfd_arbeitsunf`
FK:	`lfd_pat` -> `lei_stammdaten`, `arbeitsunf_vtr` -> `mat_kostentraeger`, `arbeitsunf_diagnose` -> `mat_icd10bmg`
Anmerkung:	entspricht Tabelle au in Rohdaten
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/190_lei_arbeitsunfaehigkeit_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/191_lei_arbeitsunfaehigkeit_transfer.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/192_lei_arbeitsunfaehigkeit_cleanIcd.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/193_lei_arbeitsunfaehigkeit_fk`

Diese Relation enthält Arbeitsunfähigkeitsmeldungen (Krankenstand) und die dazugehörigen Diagnosen.

Attribut	Typ	Verweis	Bemerkung
`lfd_arbeitsunf`	Serial	PK	Eindeutige Nummer der Arbeitsunfähigkeitsmeldung
`lfd_pat`	Integer	FK, NN	Eindeutige Kennung des Individuums
`arbeitsunf_von`	Date	NN	Beginndatum der Arbeitsunfähigkeit
`arbeitsunf_bis`	Date	NN	Enddatum der Arbeitsunfähigkeit
`arbeitsunf_diagnose`	Varchar(6)	FK	Diagnose als ICD-10 codiert
`arbeitsunf_vtr`	Varchar(2)	FK	Versicherungsträger

Überführen der Rohdaten:

Im ersten Schritt werden Datensätze zu nicht vorhandenen Pseudonymen ausgeschlossen. Diese Maßnahme betrifft lediglich 6 Individuen. Des Weiteren werden Datensätze welche z.B. das Enddatum vor dem Beginndatum ausweisen gelöscht. Ebenso wird mit duplizierten Datensätzen verfahren. Dieser Prozess ist in Abbildung 12 dargestellt.

Die Rohdaten dieser Relation enthalten zur Codierung der Diagnosen verschiedene Codesysteme (ICD-9 und ICD-10) welche in diesem Arbeitsschritt angepasst werden. Nicht jeder Eintrag in dieser Relation besitzt eine zugeordnete Diagnose. Um eine einheitliche Datenlage zu erhalten werden alle Diagnosen auf das ICD-10 BMG 2017 Codesystem (herausgegeben vom Bundesministerium für Gesundheit und Frauen in der Version BMGF-Version 2017) konvertiert (vgl. hierzu die Anpassungen in Tabelle `lei_aufenthalt_diagnose`). Die Relation enthält 2.484 Einträge mit 225 Diagnosecodes welche als ICD-9 codiert sind. Dies entspricht etwa 2,5% der vorhandenen Datensätze in der Relation `lei_arbeitsunfaehigkeit`.

Ein Konvertieren mittels der dafür vorgesehen Tabelle der GAP-DRG2 scheitert an fehlenden und mehrdeutig beschriebenen Codes. Die Daten werden in weiterer Folge mithilfe der Daten aus der Datei SYSTEMDATEN2013.accdb (siehe owncloud@dexhelpp: Material/GAP-DRG/GE-Metadaten-All-Original/Casemix2/SYSTEMDATEN2013.accdb) übersetzt. Durch diesen Schritt können 170 der 225 Codes übersetzt werden. Im folgenden Arbeitsschritt werden alle übrigen Codes über die Webseite des OHDSI Athena[3] Projekts in ein SNOMED CT Konzept übersetzt und anschließend versucht den entsprechenden ICD-10 Code zu erhalten. Nach diesem Arbeitsschritt verbleiben noch 40 ungeklärte Codes, welche in weiterer Folge manuell in die ICD-10 BMG 2017 Systematik übersetzt werden. Hierzu wird die Webseite des Deutschen Instituts für Medizinische Dokumentation und Information [5] verwendet um die Bezeichnung der ICD-9 Codes zu erhalten und anschließend wird manuell der passende ICD-10 BMG 2017 gesucht. Es verbleiben nach dem letzten Arbeitsschritt 11 nicht zuordenbare Codes. Dies betrifft 90 Datensätze in der Relation, was weniger als 0,1% der Datensätze entspricht. Aufgrund der geringen Menge wird von einer weiteren Behandlung dieser abgesehen und diese ungeklärten Codes ausgeschlossen.

Die entsprechenden Arbeitsschritte können in der Excel Datei data/lei_arbeitsunfaehigkeit_diagnose_ICD9-ICD10.xlsx eingesehen werden und werden mit dem SQL Script qi-ka/schema_modify/ 192_lei_arbeitsunfaehigkeit_cleanIcd.sql auf die Daten angewendet.

Die verbleibenden ICD-10 Codes sind in den Rohdaten ohne den separierenden Punkt notiert. Dieser Punkt wird eingefügt um einheitliche Datenformate über das gesamte Schema zu erhalten. Im Folgenden werden die in der Relation enthaltenen ICD-10 Codes soweit verändert, dass sie den gewünschten ICD-10 BMG 2017 Codes entsprechen. Bei den bearbeiteten Codes handelt es sich beispielsweise um Überkategorien welche im ICD-10 BMG 2017 nicht zur Codierung verwendet werden. Dieser Bearbeitung unterliegen 572 verschiedene Codes welche von etwa 20% der Datensätze verwendet werden. Die Bearbeitung beschränkt sich meist auf eine Zuordnung auf eine Subkategorie des ICD-10 Systems (also in den meistens Fällen auf die .9 Kategorie “nicht näher bezeichnet”). Eine Liste mit allen bearbeiteten Codes liegt in Form einer Excel Datei im Projektverzeichnis vor (siehe `DEXHELPP-GitLab: data/lei_arbeitsunfaehigkeit_diagnose_ICD10_BMG_mapping.xlsx`). Nach diesem Verarbeitungsschritt verbleiben 14 nicht konvertierte Codes welche in 103 Datensätzen

[3] http://athena.ohdsi.org/

vorkommen was in etwa 1% der Datensätze entspricht. Diese verbleibenden, nicht zuordenbaren Codes sind ebenfalls im oben genannten Excel-File einsichtig.

Die durchgeführten Anpassungen entsprechen von der Vorgehensweise derjenigen aus Tabelle `lei_aufenthalt_diagnose`. Die durchgeführten Anpassungen werden in Abbildung 12 dargestellt.

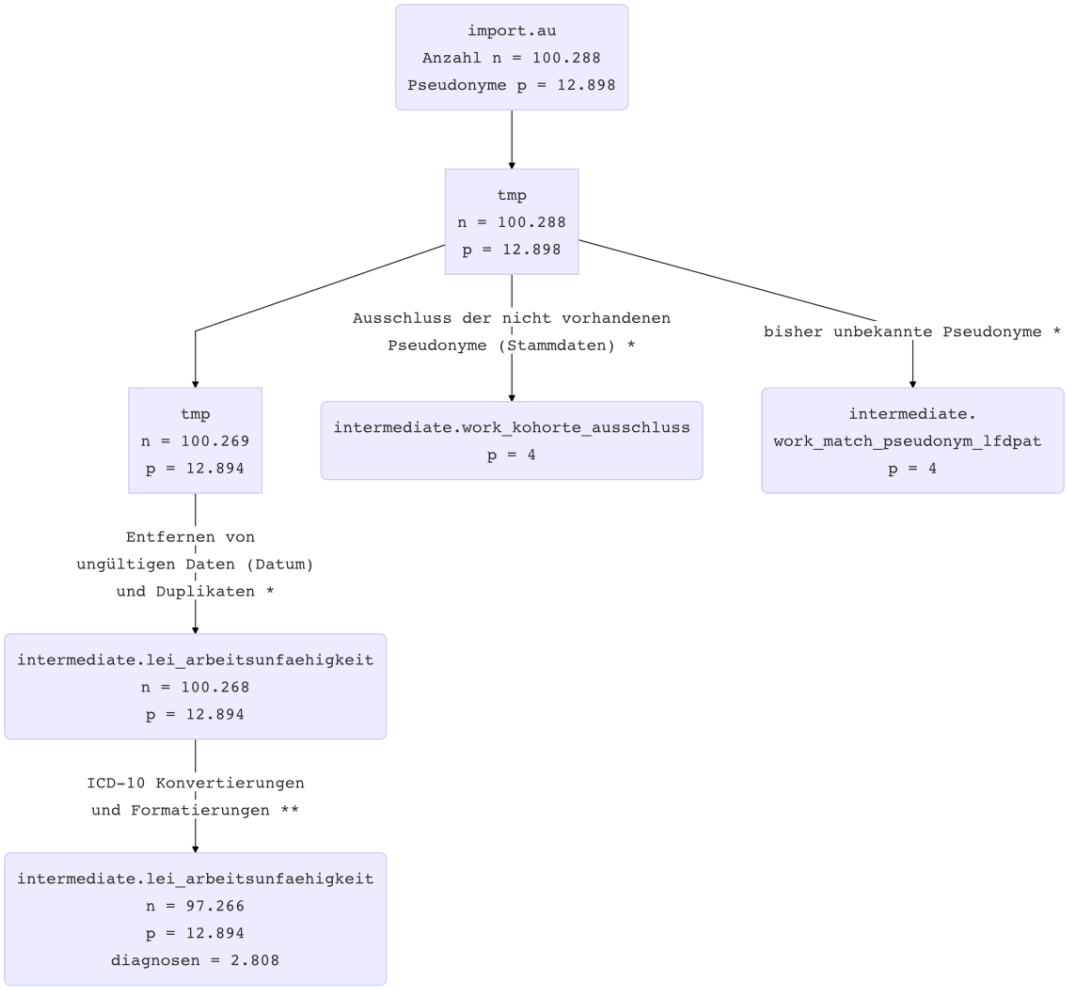

Abbildung 12: Überführung der Daten zur Arbeitsunfähigkeit
Skript 191_lei_arbeitsunfaehigkeit_transfer.sql (*) und Anpassung der Diagnosecodes in der Relation lei_arbeitsunfaehigkeit mit dem Skript 192_lei_arbeitsunfaehigkeit_cleanIcd.sql (**)

```
/* 5 random entries */
SELECT *
    FROM intermediate.lei_arbeitsunfaehigkeit
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

lfd_arbeitsunf	lfd_pat	arbeitsunf_von	arbeitsunf_bis
43167	35122	2011-10-11	2012-01-10

54104	8828	2013-02-09	2013-02-17
28133	22993	2013-06-26	2013-07-05
14624	35898	2011-09-28	2011-10-11
27469	928	2013-02-04	2013-02-08

arbeitsunf_diagnose	arbeitsunf_vtr
I20.0	18
J06.8	11
J06.9	14
A09.9	12
E10.9	14

3.5.10 Tabelle `lei_reha`

Beinhaltet Daten zu Rehaaufenthalten. Diese Daten wurden in einem späten Projektabschnitt erhalten und werden daher hier nicht in vollem Umfang aufbereitet.

Tabelle:	`lei_reha`
Beschreibung:	Beinhaltet Rehaaufenthalte
Anzahl Einträge:	7.492
PK:	`(lfd_pat, beginn)`
FK:	`lfd_pat` -> work_match_pseudonym_lfdpat
Anmerkung:	entspricht Tabelle `reha` in Rohdaten
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/210_lei_reha_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/212_lei_reha_insert.sql`

Attribut	Typ	Verweis	Bemerkung
`lfd_pat`	Integer	PK, FK	Eindeutige Kennung des Individuums
`beginn`	Date	PK	Beginndatum des Aufenthaltes
ende	Date	NN	Enddatum des Aufenthaltes
`aufnahmeart`	Varchar(1)		Aufnahmeart
`entlassungsart`	Varchar(1)		Entlassungsart

Überführen der Rohdaten:

Im ersten Schritt werden Datensätze zu nicht vorhandenen Pseudonymen ausgeschlossen. Des Weiteren werden Reha-Aufenthalte eines Patienten, welche zwar dasselbe Beginndatum, aber unterschiedliche Entlassungsdaten haben zusammengeführt da ein Patient nicht auf mehreren Rehabilitationen gleichzeitig sein kann, hierbei wird jeweils das spätere Entlassungsdatum angenommen.

Durch die erfolgte Bereinigung werden 2 duplizierte Aufenthalte als ungültig ausgeschlossen.

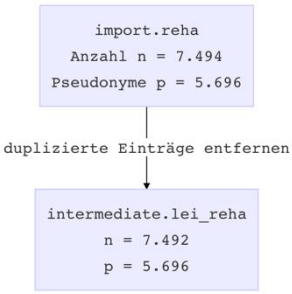

Abbildung 13: Überführung der Rehadaten

Skript 212_lei_reha_insert.sql

```
/* distinct_patient_count vs. Row_count */
SELECT COUNT(DISTINCT lfd_pat) AS distinct_patient_count,
COUNT(*) AS row_count
FROM intermediate.lei_reha;
```

distinct_patient_count	row_count
5696	7492

3.6 Material-Tabellen (mat_)

Die im Folgenden beschriebenen Relationen beinhalten Daten welche zum Zwecke der Annotation der Rohdaten eingefügt werden. Die eingefügten Daten können im Verzeichnis DEXHELPP-GITLAB:qi-ka/data/* gefunden werden, zusätzlich werden die Daten im Projekt Datenspeicher (Owncloud) bereitgestellt. Die beschriebenen Relationen sind überblicksmäßig in Abbildung 14 dargestellt.

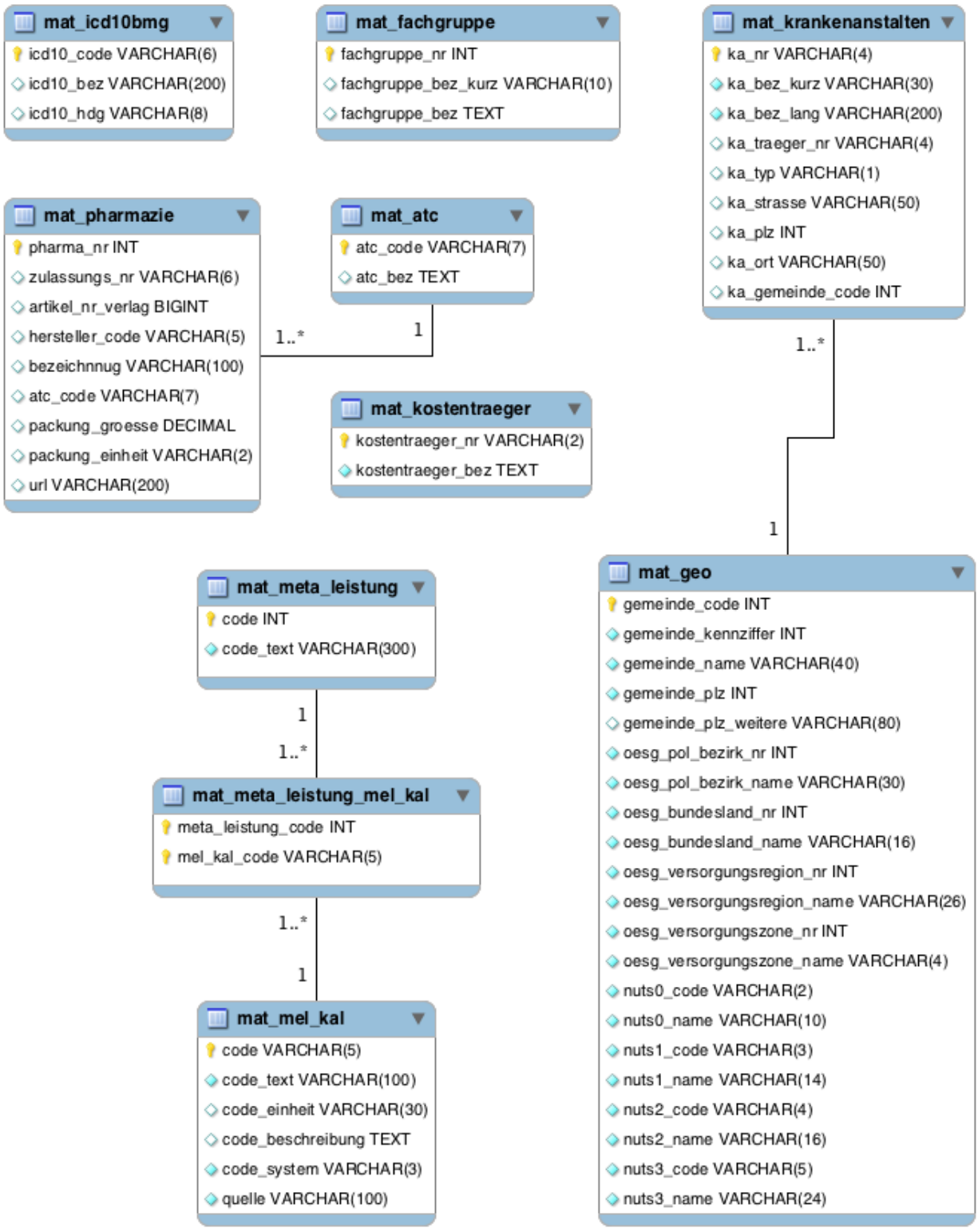

*Abbildung 14: Überblick über die Relationen mit dem Präfix mat_*

3.6.1 Tabelle `mat_krankenanstalten`

Beinhaltet alle im Datensatz vorhandenen österreichischen Krankenanstalten samt deren Bezeichung und weiteren Daten, wie zum Beispiel den Ort.

Tabelle:	mat_krankenanstalten
Beschreibung:	Beinhaltet Bezeichnungen und sonstige Attribute der Krankenanstalten
Anzahl Einträge:	316
PK:	ka_nr
FK:	ka_gemeinde_code -> mat_geo
Anmerkung:	zugehörige Daten: data/mat_krankenanstalten.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/010_mat_krankenanstalten_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/011_mat_krankenanstalten_insert.sql

Diese Tabelle beinhaltet Daten über Krankenanstalten in Österreich. Die Daten stammen aus dem GAP-DRG Projekt und wurden um die in Bezug auf das QI-KA Projekt fehlenden Krankenanstalten erweitert. Als Quelle diente das Dokument "Krankenanstalten in Österreich 2008" des Bundesministeriums für Gesundheit, Familie und Jugend, Bereich I/C und das Krankenanstalten – Online-Verzeichnis des BMFG [6].

Die Daten zu den Krankenhäusern wurden weiters um den Gemeindecode, wie in der Relation mat_geo verwendet wird, erweitert.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
ka_nr	Varchar(4)	PK	Eindeutige Kennung der Krankenanstalt
ka_bez_kurz	Varchar(30)	NN	Kurzbezeichnung der Krankenanstalt
ka_bez_lang	Varchar(200)	NN	Bezeichung der Krankenanstalt
ka_traeger_nr	Varchar(4)	-	Identifikationsnummer des Trägers
ka_typ	Varchar(1)	-	Art der Finanzierung L = Landesfonds, N = Sonstige, P = Privat
ka_strasse	Varchar(50)	-	Adresse
ka_plz	Integer	-	Postleitzahl
ka_ort	Varchar(50)	-	Stadt
ka_gemeinde_code	Integer	FK	Gemeindecode

```sql
/* 5 random entries */
SELECT *
    FROM intermediate.mat_krankenanstalten
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

ka_nr	ka_bez_kurz	ka_bez_lang	ka_traeger_nr
K812	Gaschurn SAN	Vital-Zentrum Felbermayer	T808
K624	Graz KSR PKL	Privatklinik der Kreuzschwestern	T604

K105	Kittsee LKH	Ladislaus Batthyany-Strattmann Krankenhaus Kittsee	T101
K327	Stein STRAFA	Justizanstalt Stein	T097
K828	Frastanz KH	Krankenhaus Stiftung Maria Ebene Frastanz	T822

ka_typ	ka_strasse	ka_plz	ka_ort	ka_gemeinde_code
P	Dorfstraße 20a	6793	Gaschurn	80110
P	Kreuzgasse 35	8010	Graz	60101
L	Hauptplatz 3	2421	Kittsee	10711
N	Steiner Landstraße 4	3500	Krems an der Donau	30101
L	Maria Ebene 17	6820	Frastanz	80405

```
/* krankenanstalten_count */
SELECT COUNT(*) AS krankenanstalten_count FROM intermediate.mat_krankenans
talten;
```

krankenanstalten_count
316

```
/* most common ka_traeger_nr */
SELECT ka_traeger_nr, COUNT(ka_traeger_nr)
    FROM intermediate.mat_krankenanstalten
    GROUP BY ka_traeger_nr
    ORDER BY COUNT(ka_traeger_nr) DESC
    LIMIT 10;
```

ka_traeger_nr	count
T099	21
	21
T601	20
T301	19
T090	15
T081	11
T401	11
T201	10
T801	6
T501	5

3.6.2 Tabelle `mat_kostentraeger`

Beinhaltet alle im Datensatz vorhandenen Kostenträger. Diese sind zum einen alle Versicherungsträger sowie weitere Kostenträger. Es sind gesamt 117 Kostenträger verzeichnet.

Tabelle:	`mat_kostentraeger`
Beschreibung:	Beinhaltet Bezeichnungen der verfügbaren Versicherungsträger sowie sonstiger Kostenträger

Anzahl Einträge:	117
PK:	kostentraeger_nr
FK:	-
Anmerkung:	zugehörige Daten: data/mat_kostentraeger.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/050_mat_kostentraeger_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/051_mat_kostentraeger_insert.sql

Diese Tabelle beinhaltet Versicherungsträger und sonstige Kostenträger mit ihren Codes bzw. Bezeichnungen. Die Daten dieser Relation stammen aus der GAP-DRG-2 (siehe gapdrg_datenbank_semantisches_datenmodell_mbds_2011-03-11_.pdf) bzw. Handbuch zur Dokumentation in landesgesundheitsfondsfinanzierten Krankenanstalten (Anhang 1) 2015.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
kostentraeger_nr	Varchar(2)	PK	Eindeutige Kennung des Versicherungscodes bzw. Trägercodes
kostentraeger_bez	Text	NN	Name des Kostenträgers

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_kostentraeger
    ORDER BY RANDOM() LIMIT 5;
```

kostentraeger_nr	kostentraeger_bez
49	Sozialversicherungsanstalt der gewerbl. Wirtschaft – Vorarlberg
A2	Arbeitsmarktservice Niederösterreich
8D	Kranken- und Unfallfürsorge der Tiroler Landeslehrer
M6	Bundesministerium für Innovation und Zukunft
2A	Krankenfürsorge der Beamten der Stadtgemeinde Baden

```
/* kostentraeger_count */
SELECT COUNT(*) AS kostentraeger_count FROM intermediate.mat_kostentraeger
;
```

kostentraeger_count
117

3.6.3 Tabelle mat_fachgruppe

Medizinische Fachbereiche der versorgenden Akteure welche z.B. beim Verrechnen der Leistungen angegeben werden.

Tabelle:	mat_fachgruppe
Beschreibung:	Beinhaltet Bezeichnungen der medizinischen Fachgruppen
Anzahl Einträge:	90
PK:	fachgruppe_nr
FK:	-
Anmerkung:	zugehörige Daten: data/mat_fachgruppe.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/060_mat_fachgruppe_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/061_mat_fachgruppe_insert.sql

Diese Tabelle beinhaltet Fachgruppenbezeichnungen zum Zwecke der Annotation und Prüfung vorhandener Daten, insbesondere die Daten in der Relation lei_leistung. Fachgruppen bilden die Zugehörigkeit zu einer Profession bzw. Dienstleistergruppe ab. Die Daten in dieser Relation wurden aus verschieden Quellen zusammengetragen (siehe Excel-File unter data/mat_fachgruppe.xlsx). Weiters werden zusätzlich die einzelnen Fachgruppen zu höhergeordneten Gruppen zusammengefasst um Auswertungen unabhängig von beispielsweise der Unterscheidung zwischen Wahl- und Kassenarzt durchführen zu können. Diese Vorgehensweise wird aus dem Projekt ADE-PIM übernommen, von welchem auch die Gruppeneinteilungen stammen. Neben den erklärten Daten befinden sich noch fünf weitere Fachgruppen im Datensatz, welche derzeit noch nicht beschrieben sind. Dabei handelt es sich um die Fachgruppen mit den Nummern 44, 45, 57, 76 und 79. Für eine einstweilige Validierung der Daten werden diese ungeklärten Nummern dennoch eingefügt mit dem Verweis auf fehlende Bezeichnungen.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
fachgruppe_nr	Integer	PK	Eindeutige Identifizierung der Fachgruppe
fachgruppe_bez_kurz	Varchar(10)	-	Kurzbezeichnung der Fachgruppe, falls nicht angegeben werden ersten 10 Zeichen der Bezeichnung verwendet
fachgruppe_bez	Text	-	Bezeichung der Fachgruppe
fachgruppe_gruppe_nr	Integer	-	Zuordnung von Fachgruppen zu einer übergeordneten Gruppe - Nummer dieser Gruppe. Die zweite Normalform wird hier ganz bewusst verletzt um eine gewisse Leserlichkeit der Relationen zu behalten.
fachgruppe_gruppe_bez	Varchar(300)	-	Zuordnung von Fachgruppen zu einer übergeordneten Gruppe - Bezeichnung dieser Gruppe

```
/* 5 random entries */
SELECT *
```

```
    FROM intermediate.mat_fachgruppe
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

fachgruppe_nr	fachgruppe_bez_kurz	fachgruppe_bez
33	Hist	FA für Histologie und Embryologie
35	Vir	FA für Virologie
92	HKraP	Hauskrankenpflegeorganisation
96	AmbKFO	Selbständiges Ambulatorium für Kieferorthopädie
16	Uro	FA für Urologie

fachgruppe_gruppe_nr	fachgruppe_gruppe_bez
0	FGB wurde in Kollektiv nicht verrechnet, daher keine Gruppenzuordnung
0	FGB wurde in Kollektiv nicht verrechnet, daher keine Gruppenzuordnung
0	FGB wurde in Kollektiv nicht verrechnet, daher keine Gruppenzuordnung
2	Gebiete der Chirurgie, Unfallchirurgie und Orthopädie
7	Gebiete der Haut- und Geschlechtskrankheiten und der Urologie

3.6.4 Tabelle `mat_icd10bmg`

Diagnosecodes lt. ICD-10 des Bundesministeriums für Gesundheit und Frauen (BMG) in der Version 2017 [4].

Tabelle:	`mat_icd10bmg`
Beschreibung:	Beinhaltet ICD-10 Diagnosecodes lt. Bundesministerium für Gesundheit und Frauen
Anzahl Einträge:	13.184
PK:	`icd10_code`
FK:	-
Anmerkung:	-
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/070_mat_icd10bmg_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/071_mat_icd10bmg_insert.sql`

Diese Relation enthält ICD-10 Codes entsprechend dem Bundesministerium für Gesundheit und Frauen (BMG). Vergleiche die hierzu bereitgestellten Dokumente der

Onlineressourcen des BMG. Zur Verwendung kommt die Version ICD-10-BMG 2017. Die Version 2018 ist zum Zeitpunkt der Projektabwicklung zwar zugänglich, jedoch enthalten die Rohdaten nur Datensätze bis ins Jahr 2017.

Neben den ICD-10 Codes welche durch die World Health Organization (WHO) herausgegeben werden enthält die BMG Version zusätzliche nummerische Codes welche für besondere Ereignisse verwendet werden.

Die eingefügten Diagnosen sind unter `DEXHELPP-GITLAB: qi-ka/data/ICD-10_BMG_src/*` bzw. unter `DEXHELPP-GITLAB: qi-ka/data/mat_icd10bmg.xlsx` zu finden.

Attribut	Typ	Verweis	Bemerkung
`icd10_code`	Varchar(6)	PK	ICD-10 Code lt. Definition in der Form X00.00 oder X00.0
`icd10_bez`	Varchar(200)	-	Beschreibung der Diagnose in deutscher Sprache
`icd10_hdg`	Varchar(8)	-	zugerechnete Hauptdiagnosegruppe
`icd10_kapitel`	Varchar(5)	-	ICD-10 Kapitel entsprechend des ICD-10 BMG 2017

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_icd10bmg
    ORDER BY RANDOM() LIMIT 5;
```

icd10_code	icd10_bez	icd10_hdg	Icd10_kapitel
M43.90	Deformität der Wirbelsäule und des Rückens, nicht näher bezeichnet: Mehrere Lokalisationen der Wirbelsäule	HDG01.32	XIII
M89.51	Osteolyse: Schulterregion	HDG14.07	XIII
N85.6	Intrauterine Synechien	HDG11.08	XIV
C67.9	Bösartige Neubildung: Harnblase, nicht näher bezeichnet	HDG09.02	II
R50.9	Fieber, nicht näher bezeichnet	HDG16.11	XVIII

3.6.5 Tabelle `mat_pharmazie`

Medikamente lt. Apothekerverlag samt zugeordneter Stoffgruppencodes (ATC).

Tabelle:	`mat_pharmazie`
Beschreibung:	Beinhaltet Informationen zu Medikamenten, großteils sind dies Produkte des Warenverzeichnisses des Apothekerverlages
Anzahl Einträge:	11.250
PK:	`pharma_nr`
FK:	`atc_code` -> `mat_atc`
Anmerkung:	zugehörige Daten: `data/mat_pharmazie.xlsx` (alle im Projekt relevante Codes) bzw. den Dateien im Ordner

`data/mat_pharmazie/*` (enthält Daten aus verschiedenen Quellen und deren Aufbereitung)

SQL Scripts: `DEXHELPP-GITLAB: qi-ka/schema_modify/090_mat_pharmazie_create.sql`
`DEXHELPP-GITLAB: qi-ka/schema_modify/091_mat_pharmazie_insert.sql`

Diese Tabelle beinhaltet Produkte, deren Eigenschaften und zugeordnete ATC Codes des Warenverzeichnisses des österreichischen Apotheker Verlages. Die Daten stammen aus dem Arzneimittelverzeichnis des Apotheker Verlages. Einige, im Arzneimittelverzeichnis nicht vorkommende Produkte, wurden durch die Daten der Projekte ADE-PIM und GAP-DRG2 ergänzt. Alle Produkte aus dem Arzneimittelverzeichnis, welche ausschließlich für den veterinärmedizinischen Einsatz gedacht sind, wurden ausgeschlossen.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
pharma_nr	Integer	PK	Pharmazentralnummer des Präparates
zulassungs_nr	Varchar(6)	-	Nummer der Zulassung des Arzneimittels
artikel_nr_verlag	Bigint	-	Interne Artikelnummer des Präparates des Apotheker Verlages
hersteller_code	Varchar(5)	-	Codierung des Arzneimittelherstellers, Codesystem lt. Apotheker Verlag
bezeichnung	Varchar(100)	NN	Bezeichnung des Arzneimittels bzw. sonstigen Produktes
atc_code	Varchar(7)	FK	ATC Code lt. WHO
packung_groesse	Numeric	-	Mengenangabe
packung_einheit	Varchar(2)	-	Mengeneinheiten
url	Varchar(200)	-	optionale URL zum Hersteller

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_pharmazie
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

pharma_nr	zulassungs_nr	artikel_nr_verlag	hersteller_code
2309256	-	5318900812	RAU02
4220879	136626	523506120	GEN01
1141122	118732	661852000	PFI01
1265829	119689	1336204000	MER03
1338103	-	-	-

bezeichnnug	atc_code	packung_groesse	packung_einheit	url
MULLK.GAZIN ST 8F 7,5X7,5	-	100	ST	-

DULOXETIN Genericon 60 mg - magensaftresistente Hartkapseln	N06AX21	30	ST	-
FRAGMIN 10 000 IE/1 ml - Ampullen	B01AB04	10	ST	-
PROSCAR 5 mg - Filmtabletten	G04CB01	28	ST	-
Klacid 500 mg Filmtabl.	J01FA09	-	-	-

3.6.6 Tabelle `mat_atc`

Stoffgruppencodes lt. dem Anatomical Therapeutic Chemical-Codes der WHO in der aktuellen Fassung (2018-08)

Tabelle:	`mat_atc`
Beschreibung:	Beinhaltet ATC-Codes und deren Bezeichnung
Anzahl Einträge:	6.205
PK:	`atc_code`
FK:	-
Anmerkung:	zugehörige Daten: `data/mat_atc.csv`
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/080_mat_atc_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/081_mat_atc_insert.sql`

Diese Tabelle beinhaltet Anatomical Therapeutic Chemical-Codes der WHO, in deutscher Sprache. Die Daten stammen aus dem Arzneimittelverzeichnis des Apotheker Verlages. Die enthaltenen Daten stellen keinen Anspruch auf Vollständigkeit, vielmehr sind alle Codes enthalten welche in den Rohdaten vorkommen und einige mehr (jedoch ausschließlich der Gruppe 'Q - Veterinärmedizinische Arzneimittel').

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
`atc_code`	Varchar(7)	PK	Eindeutige Kennung entsprechend der ATC Systematik
`atc_bez`	Text	-	Beschreibung des Codes in deutscher Sprache

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_atc
    ORDER BY RANDOM() LIMIT 5;
```

atc_code	atc_bez
J06AA04	Botulismus-Antitoxin
A08AA08	Clobenzorex
A02AB03	Aluminiumphosphat

R06AD03	Thiethylperazin
A02AD02	Magaldrat

3.6.7 Tabelle `mat_mel_kal`

Codes zu medizinischen Einzelleistungen und dem Katalog ambulanter Leistungen, entsprechend dem Leistungskatalog 2018 des Bundesministeriums für Gesundheit und Frauen [3].

Tabelle:	`mat_mel_kal`
Beschreibung:	Beinhaltet MEL (Medizinische Einzelleistungen) und KAL (Katalog ambulanter Leistungen) Codes
Anzahl Einträge:	2.102
PK:	`code`
FK:	-
Anmerkung:	zugehörige Daten: `data/mat_mel-kal.xlsx`
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/095_mat_mel_kal_create.sql` `DEXHELPP-GITLAB: qi-ka/schema_modify/095_mat_mel_kal_insert.sql`

Diese Tabelle beinhaltet MEL-Codes und KAL-Codes aus verschiedenen Dokumenten des BMG und dient der Annotation der vorhandenen Daten. Die Rohdaten können im Excel File wie oben benannt, samt deren Quellen eingesehen werden. Als Grundlage wurde der Leistungskatalog 2017 des BMGF genommen und fehlende Codes (aus den vorigen Jahren, etc.) ergänzt. Zusätzlich wurden ca. 20 Codes aus dem neuen Leistungskatalog 2018 des BMGF eingefügt.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
`code`	Varchar(5)	PK	MEL oder KAL Code
`code_text`	Varchar(100)	NN	Bezeichnung des MEL oder KAL Codes
`code_einheit`	Varchar(30)	-	Beschreibt in welcher Einheit dieser Code anzuwenden ist (z.B. pro Behandlung)
`code_beschreibung`	Text	-	Detailliertere Beschreibung des Codes
`code_system`	Varchar(3)	NN	Bezeichnung des Codesystems (‘MEL’ oder ‘KAL’)
`quelle`	Varchar(100)	NN	Angabe woher der Code stammt (welches Dokument, Jahr)

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_mel_kal
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

code	code_text	code_einheit
ZN600	Sonographie von oberflächlichen Raumforderungen (LE=je Sitzung)	je Sitzung
MD030	Osteosynthese mehrerer Mittelhandknochen, Finger (LE=je Seite)	je Seite
AK050	Thorakale oder lumbale Sympathektomie – thorakoskopisch/laparoskopisch (LE=je Sitzung)	je Sitzung
XA076	Onkologische Therapie – monoklonaler Antikörper Obinutuzumab (LE=je Applikation)	je Applikation/Prothese/Stent
AG040	Kraniozervikale Erweiterungsplastik der Dura (LE=je Sitzung)	je Sitzung

code_beschreibung	code_system	quelle
Ultraschalluntersuchung von oberflächlichen Raumforderungen (z.B. Zysten, Tumore, Hämatome)	KAL	Leistungskatalog BMGF 2017 - Tabelle SP2 (gesamt) (20.03.2017)
-	MEL	Leistungskatalog BMGF 2017 - Tabelle SP2 (gesamt) (20.03.2017)
-	MEL	Leistungskatalog BMGF 2017 - Tabelle SP2 (gesamt) (20.03.2017)
Entsprechend 1000 mg, Aufteilung auf 100mg + 900mg entspricht einer Applikation	MEL	Leistungskatalog BMGF 2017 - Tabelle SP2 (gesamt) (20.03.2017)
-	MEL	Leistungskatalog BMGF 2017 - Tabelle SP2 (gesamt) (20.03.2017)

3.6.8 Tabelle `mat_meta_leistung`

Abstrakte Leistungen aus dem niedergelassenen Bereich um Auswertungen der Leistungen über mehrere Versicherungsträger (und deren verschiedenen Notationen) hinweg zu erlauben.

Tabelle:	mat_meta_leistung
Beschreibung:	Beinhaltet Metaleistungen aus dem niedergelassenen Bereich
Anzahl Einträge:	1.683
PK:	code
FK:	-
Anmerkung:	zugehörige Daten: data/mat_meta_leistung/mat_meta_leistung.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/093_mat_meta_leistung_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/094_mat_meta_leistung_insert.sql

Diese Tabelle enthält abstrakte Leistungen welche ein Metamodell zu den Leistungen der einzelnen Versicherungsträgern darstellt. Die Leistungen der Versicherungsträger werden teils verschieden granular und teils in unterschiedlichen Notationen codiert. Die Daten aus dieser Relation stammen aus einer Anfrage bei der NÖGKK und stellen den aktuellen Stand (2018-08-03) des Hono Datenbestandes seitens der Niederösterreichischen Gebietskrankenkassa dar. Die Tabelle wird im Rahmen des Projektes verwendet um den einzelnen Leistungen einen MEL-Code bzw. KAL-Code zuzuordnen und zwar unabhängig von dem jeweiligen Versicherungsträger.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
code	Int	PK	interner Code der Metaleistung
code_text	Varchar(300)	NN	textuelle Beschreibung der Leistung

3.6.9 Tabelle mat_meta_leistung_mel_kal

Zuordnung der zuvor eingeführten Metaleistungen zu medizinischen Einzelleistungen und Leistungen aus dem Katalog ambulanter Leistungen lt. Leistungskatalog 2018 des Bundesministeriums für Gesundheit und Frauen.

Tabelle:	mat_meta_leistung_mel_kal
Beschreibung:	Beinhaltet eine Zuordnung von MEL bzw. KAL Codes zu den Metaleistungen aus dem niedergelassenen Bereich
Anzahl Einträge:	1.967
PK:	meta_leistung_code, mel_kal_code
FK:	meta_leistung_code -> mat_meta_leistung, mel_kal_code -> mat_mel_kal
Anmerkung:	zugehörige Daten: data/mat_meta_leistung/mat_meta_leistung.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/097_mat_meta_leistung_mel_kal_create.sql DEXHELPP-GITLAB:

qi-ka/schema_modify/098_mat_meta_leistung_mel_kal_insert.sql

Diese Tabelle ordnet den Metaleistungen einen oder mehrere MEL bzw. KAL Code zu. Die Daten aus dieser Relation stammen ebenfalls aus einer Anfrage bei der NÖGKK (Ansprechperson: Hr. Rohbausch) und stellen den aktuellen Stand (2018-08-03) des Hono Datenbestandes seitens der Niederösterreichischen Gebietskrankenkassa dar.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
meta_leistung_code	Int	PK, FK, NN	interner Code der Metaleistung
mel_kal_code	Varchar(5)	PK, FK, NN	MEL oder KAL Code

3.6.10 Tabelle mat_geo

Geographische Einteilung von österreichischen Gemeinden und deren Zugehörigkeit zu den, seitens der europäischen Union, definierten NUTS-Regionen, laut Statistik Austria [2]. Zusätzlich wird die Zugehörigkeit zu den Versorgungsregionen und -zonen des österreichischen Strukturplans Gesundheit [7], [8] abgebildet.

Tabelle:	mat_geo
Beschreibung:	Beinhaltet Daten zu Orten und deren Zugehörigkeit zu NUTS Regionen bzw. Versorgungsregionen und Versorgungszonen des ÖSG (Österreichischer Strukturplan Gesundheit)
Anzahl Einträge:	2.120
PK:	gemeinde_code
FK:	-
Anmerkung:	zugehörige Daten: data/regionen/mat_geo.xlsx
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/005_mat_geo_create.sql DEXHELPP-GITLAB: qi-ka/schema_modify/006_mat_geo_insert.sql

Die Daten stammen aus verschiedenen Quellen und sind im oben genannten Excel File beschrieben. Als Vorlage für diesen Prozess diente das Projekt ADE3 und das zugehörige Dokument ‘Projekt ADE3 – Endbericht’ (im selben Ordner wie die oben genannten Daten zu finden).

Die Tabelle beinhaltet lt. Gebietsstand zum 1.1.2018 der Statistik Austria alle Gemeinden Österreichs. Diese wurden entsprechend der Zuordnung (ebenfalls der Statistik Austria) zu den europäischen NUTS (Nomenclature des unités territoriales statistiques) Einheiten annotiert. Die Einordnung der NUTS Regionen wird von Level 0 (Nationalstaaten) bis Level 3 (Zusammenfassung von mehreren Gemeinden) durchgeführt.

Des Weiteren wurden die Gemeinden den Versorgungsregionen und Versorgungszonen des ÖSGs (Österreichischer Strukturplan Gesundheit) zugeordnet, entsprechend der definierten Einteilung.

Die Daten dieser Tabelle werden zur Annotation, Validierung und später zur Darstellung benötigt.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
gemeinde_code	Integer	NN	Gemeindecode (GCD) lt. Adressregisterverordnung, eindeutig für jede Gemeinde österreichweit, siehe Gemeindelisten [2]
gemeinde_kennziffer	Integer	NN	Dem Gemeindecode gleich, ausgenommen in Wien, hier ist nur eine Kennziffer für ganz Wien vergeben (90001), siehe Gemeindelisten [2]
gemeinde_name	Varchar(40)	NN	Name der Gemeinde
gemeinde_plz	Integer	NN	Postleitzahl des Gemeindeamtes
gemeinde_plz_weitere	Varchar(80)		optional weitere Postleitzahlen der Gemeinde, Textuell durch Leerzeichen getrennt
oesg_pol_bezirk_nr	Integer	NN	Nummer des politischen Bezirkes, siehe ÖSG [8]
oesg_pol_bezirk_name	Varchar(30)	NN	Bezeichnung des politischen Bezirkes
oesg_bundesland_nr	Integer	NN	Nummer des Bundeslandes, siehe ÖSG [8]
oesg_bundesland_name	Varchar(16)	NN	Name des Bundeslandes
oesg_versorgungsregion_nr	Integer	NN	Nummer der Versorgungsregion, es existieren 32 Versorgungsregionen denen jeweils ganze Bezirke zugeordnet werden, siehe ÖSG [8]
oesg_versorgungsregion_name	Varchar(26)	NN	Name der Versorgungsregion
oesg_versorgungszone_nr	Integer	NN	Nummer der Versorgungszone, es existieren 4 Versorgungszonen im ÖSG
oesg_versorgungszone_name	Varchar(4)	NN	Name der Versorgungszone
nuts0_code	Varchar(2)	NN	Code der NUTS-0 Einheit, Die Ebene NUTS 0

			entspricht dem Mitgliedsstaat der EU, siehe NUTS-Regionen [9]
nuts0_name	Varchar(10)	NN	Name der NUTS-0 Einheit
nuts1_code	Varchar(3)	NN	Code der NUTS-1 Einheit, Aufteilung Österreichs in drei Einheiten, siehe [9]
nuts1_name	Varchar(14)	NN	Name der NUTS-1 Einheit
nuts2_code	Varchar(4)	NN	Code der NUTS-2 Einheit, dies entspricht in Österreich den Bundesländern, siehe [9]
nuts2_name	Varchar(16)	NN	Name der NUTS-2 Einheit
nuts3_code	Varchar(5)	NN	Code der NUTS-3 Einheit, stellt eine Zusammenfassung von mehreren Gemeinden dar, jede Gemeinde ist genau einer NUTS-Einheit zugeordnet. Wien bildet eine eigene NUTS 3-Einheit, siehe [9]
nuts3_name	Varchar(24)	NN	Name der NUTS-3 Einheit

```
/* 5 random entries */
SELECT *
    FROM intermediate.mat_geo
    ORDER BY RANDOM() LIMIT 5;
```

Table continues below

gemeinde_code	gemeinde_kennziffer	gemeinde_name	gemeinde_plz
31617	31617	Großkrut	2143
41750	41750	Wolfsegg am Hausruck	4902
10809	10809	Lockenhaus	7442
31551	31551	Texingtal	3242
31311	31311	Gföhl	3542

Table continues below

gemeinde_plz_weitere	oesg_pol_bezirk_nr	oesg_pol_bezirk_name
-	316	Mistelbach
-	417	Vöcklabruck
-	108	Oberpullendorf
3241	315	Melk
3521 3553	313	Krems (Land)

Table continues below

oesg_bundesland_nr	oesg_bundesland_name	oesg_versorgungsregion_nr

3	Niederösterreich	33
4	Oberösterreich	45
1	Burgenland	11
3	Niederösterreich	35
3	Niederösterreich	31

Table continues below

oesg_versorgungsregion_name	oesg_versorgungszone_nr
Weinviertel	1
Traunviertel-Salzkammergut	3
Burgenland-Nord	1
Mostviertel	1
NÖ Mitte	1

Table continues below

oesg_versorgungszone_name	nuts0_code	nuts0_name	nuts1_code
Ost	AT	Österreich	AT1
Nord	AT	Österreich	AT3
Ost	AT	Österreich	AT1
Ost	AT	Österreich	AT1
Ost	AT	Österreich	AT1

Table continues below

nuts1_name	nuts2_code	nuts2_name	nuts3_code
Ostösterreich	AT12	Niederösterreich	AT125
Westösterreich	AT31	Oberösterreich	AT315
Ostösterreich	AT11	Burgenland	AT111
Ostösterreich	AT12	Niederösterreich	AT121
Ostösterreich	AT12	Niederösterreich	AT124

nuts3_name
Weinviertel
Traunviertel
Mittelburgenland
Mostviertel-Eisenwurzen
Waldviertel

3.7 Arbeits-Tabellen (work_)

Die folgenden Tabellen werden für die Projektabwicklung benötigt. Wesentliche Elemente sind die Zuordnung von Pseudonymen zu einer laufenden Patientennummer und der Ausschluss aus der Forschungspopulation mit jeweils eigenen Tabellen. Die beschriebenen Relationen sind überblicksmäßig in Abbildung 15 dargestellt.

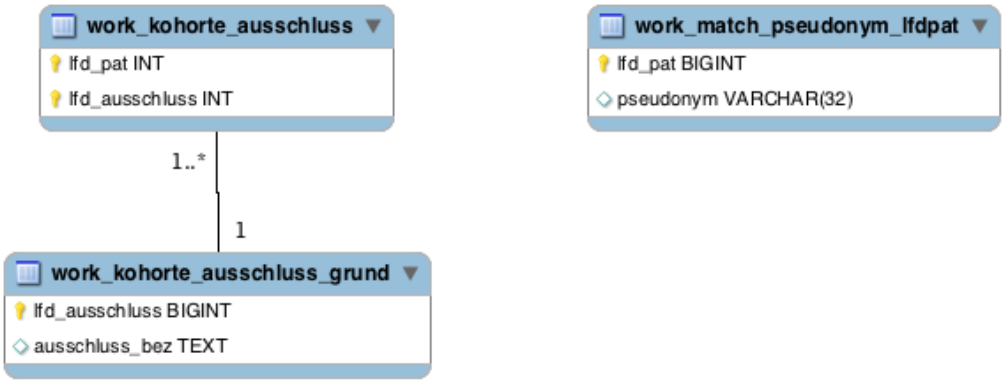

*Abbildung 15: Überblick über die Relationen mit dem Präfix work_*

3.7.1 Tabelle `work_match_pseudonym_lfdpat`

Zuordnung von Pseudonymen, welche in den Rohdaten vorkommen, zu einer Laufnummer welche ein eindeutiges Individuum im gesamten Datenbankschema beschreibt. Dies dient dem Zwecke der einfacheren Lesbarkeit der übrigen Relationen und daraus resultierenden Auswertungen.

Tabelle:	`work_match_pseudonym_lfdpat`
Beschreibung:	Zuordnung von Pseudonymen zu einer Laufnummer
Anzahl Einträge:	47.166
PK:	`lfd_pat`
FK:	-
Anmerkung:	
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/020_work_match_pseudonym_lfdpat_create.sql`

Alle im Rohdatensatz vorkommende Pseudonyme werden hier eingetragen auch wenn deren Datensatz nicht vollständig ist (nicht in allen Relationen vorkommt). Die Entscheidung eine solche Tabelle einzuführen wurde im Rahmen der Projektbesprechung 2018-06-27 getroffen.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
`lfd_pat`	Serial	PK	Eindeutige Kennung des Patienten (Laufnummer, generiert)
`pseudonym`	Varchar(32)	unique	Pseudonym aus den Originaldatensatz

3.7.2 Tabelle `work_kohorte_ausschluss`

Ausschluss von Individuen aus der beobachteten Kohorte aufgrund näher beschriebener Begründung. Auf diese Weise werden gesamt 2.818 Individuen ausgeschlossen.

Tabelle:	work_kohorte_ausschluss
Beschreibung:	Ausschluss von Pseudonymen aus der Forschungspopulation
Anzahl Einträge:	2.818
PK:	lfd_pat, ausschluss_grund_id
FK:	ausschluss_grund_id -> work_kohorte_ausschluss_grund
Anmerkung:	
SQL Scripts:	DEXHELPP-GITLAB: qi-ka/schema_modify/040_work_kohorte_ausschluss_create.sql

Diese Tabelle definiert in Verbindung mit der Tabelle work_kohorte_ausschluss_grund die Forschungspopulation indem ausgeschlossene Individuen (identifiziert durch die lfd_pat) in diese Tabelle eingetragen werden. Zusätzlich wird ein Grund für den Ausschluss angegeben. Zu einem Pseudonym können auf diese Weise mehr als ein Eintrag angelegt werden, falls mehrere Gründe für einen Ausschluss existieren.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
lfd_pat	Integer	PK	Eindeutige Kennung des Individuums
lfd_ausschluss	Integer	PK, FK -> work_kohorte_ausschluss_grund	Begründungscode des Ausschlusses siehe folgende Tabelle

```
/* 5 random entries */
SELECT *
    FROM intermediate.work_kohorte_ausschluss
    ORDER BY RANDOM() LIMIT 5;
```

lfd_pat	*lfd_ausschluss*
46471	5
46246	4
46762	6
46517	6
47077	7

```
/* row count */
SELECT COUNT(*) AS kohorte_ausschluss_count FROM intermediate.work_kohorte
_ausschluss;
```

kohorte_ausschluss_count
2818

3.7.3 Tabelle work_kohorte_ausschluss_grund

Begründung für einen Ausschluss aus der Forschungskohorte.

Tabelle:	work_kohorte_ausschluss_grund

Beschreibung:	Begründung für den Ausschluss von Individuen aus der Forschungspopulation
Anzahl Einträge:	10
PK:	`lfd_ausschluss`
FK:	-
Anmerkung:	-
SQL Scripts:	`DEXHELPP-GITLAB: qi-ka/schema_modify/030_work_kohorte_ausschluss_grund_create.sql`

Diese Tabelle beinhaltet Begründungen für einen Ausschluss aus der Forschungspopulation. Ausgeschlossene Individuen werden in die Tabelle `work_kohorte_ausschluss_grund` gespeichert und es existiert eine Beziehung zu dieser Tabelle. Die Begründungen werden in textueller Form gespeichert.

Schema der Tabelle:

Attribut	Typ	Verweis	Bemerkung
`lfd_ausschluss`	Serial	PK	Eindeutige Kennung des Ausschlussgrundes
`ausschluss_bez`	Text	-	textuelle Beschreibung des Ausschlussgrundes

```
/* 5 random entries */
SELECT *
    FROM intermediate.work_kohorte_ausschluss_grund
    ORDER BY RANDOM() LIMIT 5;
```

lfd_ausschluss	ausschluss_bez
8	Keine Stammdaten zu diesem Pseudonym vorhanden (bei import Heilmittel)
9	Keine Stammdaten zu diesem Pseudonym vorhanden (bei import Einzelkosten/Leistung)
3	Nicht einheitliche Todesdaten in Tabelle “stammdaten” der Rohdaten
2	Nicht einheitliche Geburtsdaten in Tabelle “stammdaten” der Rohdaten
10	Keine Stammdaten zu diesem Pseudonym vorhanden (bei import Arbeitsunfähigkeit)

```
/* row count */
SELECT COUNT(*) AS kohorte_ausschluss_grund_count FROM intermediate.work_k
ohorte_ausschluss_grund;
```

kohorte_ausschluss_grund_count
10

4 Plausibilitätsprüfungen und fehlende Daten

Wie in den vorhergehenden Kapiteln beschrieben wurden bereits Ausschlüsse von fehlerhaften Daten im Zuge der Migration der Rohdaten in das Schema intermediate durchgeführt. Im Folgenden werden weitere möglicherweise fehlerhafte Daten dokumentiert und die Entscheidungen wie mit denen umzugehen ist dokumentiert.

4.1 MBDS - Leicon Matching

Die Daten aus dem MBDS (Minimum Basic Data Set) des Bundesministeriums für Gesundheit und Frauen (Sozialministerium) werden in anonymisierter Form bereitgestellt. Um die Daten mit denen des Leicon Datensatzes zusammenzuführen wurde ein probabilistisches Matchingverfahren (seitens des Auftraggebers) angewendet. Da mangels fehlender Informationen über dieses Matchingverfahren keine Möglichkeit für eine Evaluation dieses Verfahrens besteht, werden in diesem Abschnitt einfache Plausibilitätsprüfungen zwischen den MBDS Daten und den restlichen vorhandenen Daten durchgeführt. Beispielsweise wird, sollte sich das Geschlecht oder Alter eines Patienten zwischen den beiden Datenquellen unterscheiden, davon ausgegangen dass dieses Matching in diesem Fall nicht korrekt ist.

4.1.1 Inkonsistente Geschlechter

Da in beiden Datenquellen (Leicon & MBDS) ein Geschlecht enthalten ist, lassen sich fehlerhaft zugeordnete Daten dadurch eventuell identifizieren. Die entsprechenden SQL Kommandos zu dieser Fragestellung sind unter `sql_playground/2018-08-22_Plausibilitaetspruefungen.sql` zu finden. Die Attribute `pat_geschlecht` der beiden Tabellen `intermediate.lei_stammdaten` und `intermediate.lei_aufenthalt` werden miteinander verglichen und fehlerhafte Zuordnungen analysiert.

Dies betrifft 1.779 der 46.231 Patienten in der Relation `intermediate.lei_aufenthalt` und es sind dadurch 5.074 der 239.907 Datensätze in dieser Relation betroffen.

4.1.2 Inkonsistente Altersangaben

In gleicher Weise werden die Daten aus den beiden Datenquellen hinsichtlich ihres angegeben Alters (Attribut ‘pat_alter_entlassung’ & Attribut ‘aufenthalt_bis’) bzw. dem eingetragenen Geburtsdatum untersucht. Aus den Angaben im Aufenthalt wurden zu erwartende Geburtsdaten errechnet und mit den Einträgen in den Stammdaten verglichen.

Die Berechnung einer Abweichung in Jahren dieser beiden Werte ergibt folgende Verteilung:

Abweichung in Jahren	Anzahl der Patienten
0 <= x < 0,5	30.199
0,5 <= x < 1	30.305
1 < x < 2	8.981
2 <= x < 4	15.216
4 <= x < 8	9.121
8 <= x < 16	502

16 <= x < 32	481
32 <= x < 64	481
64 <= x < 128	49

Des Weiteren konnte eine Häufung dieser Abweichungen in den Jahren 2015 und 2016 festgestellt werden. Es konnte allerdings kein Zusammenhang mit dem Ort der Datenerhebung oder weiteren Attributen in der Relation hergestellt werden.

Bei einer explorativen Untersuchung zweier Patienten, bei der versucht wurde die Datensätze mithilfe der GAP-DRG2 nachzuvollziehen wurden zwei Patienten aus Niederösterreich aus dem Jahr 2011 beobachtet:

Der erste Patient (lfd_pat:13634 = pseudonym:Ca61ry7FXc7aUZuS/v+u/fJvyOdVo5tr) hat bei einem der Aufenthalte eine auffällig große Abweichung zum Geburtsdatum der Stammdaten (>3 Jahre) und wurde daher mit der GAP-DRG verglichen. Dort sind alle Aufenthalte der gleichen Person zugeordnet, außer ebendieser mit den ausreißenden Altersangaben. Zudem stimmt das Geschlecht nicht überein.

Beim zweiten nachverfolgten Patient (lfd_pat:23346 = Pseudonym:o1ZeMHo+Yqe8c4J+YhZ5IaRebg7K7frx) wurden zwar alle 10 diesem Pseudonym zugeordneten Spitalsaufenthalte in der GAP-DRG gefunden, allerdings wurden dort (GAP-DRG) der Person nur drei dieser Aufenthalte zugeordnet. Bei diesen stimmt das Alter auch überein. Von den übrigen sieben Aufenthalten war keiner einer Person zugeordnet.

4.2 Fehlende semantische Beschreibung der Rohdaten

Die semantische Beschreibung der Rohdaten erfolgte ohne Informationen über die enthaltenen Daten oder Attribute. Die meisten der Attribute konnten aufgrund der enthaltenen Daten verstanden werden. Bei einigen jedoch gelang dies trotz großer Bemühungen nicht. Im Folgenden werden alle Attribute der Rohdaten aufgelistet deren Bedeutung nicht eindeutig geklärt werden konnte:

Relation	Attribut	Beschreibung
mbds_sa1_aufenthalte	pstidtyp	numerischer Wert, enthaltene Werte nur NULL oder 3. Es wird vermutet, dass es sich hierbei um ein Maß für die Anonymität des Datensatzes handelt, dies ist jedoch reine Spekulation.
Mbds_sa1_aufenthalte	zeitraum	Datumswert, 40 verschiedene Werte, >50% fehlende Daten (NULL). Unklare Bedeutung.
Mbds_sa1_aufenthalte	anz	numerischer Wert mit Ausprägungen NULL (~55%) oder 1 (~45%). Unklare Bedeutung.
Mbds_sa2_diagnose	zeitraum	Datumswert, 40 verschiedene Werte, >50% fehlende Daten (NULL). Unklare Bedeutung.
Mbds_sa2_diagnose	anz	numerischer Wert mit Ausprägungen NULL (~55%) oder 1 (~45%). Unklare Bedeutung.
Mbds_sa3_mel	zeitraum	Datumswert, 40 verschiedene Werte, >63% fehlende Daten (NULL). Unklare Bedeutung.

Mbds_sa3_mel	anz	numerischer Wert mit Ausprägungen NULL (~63%) oder 1 (~37%). Unklare Bedeutung.
Mbds_sa3_mel	anz2	numerischer Wert mit Ausprägungen NULL (~97%) oder 1 (~3%). Unklare Bedeutung.
Ek	honoid	scheint eine Zuordnung zu abstrahierten Leistungspositionen zu sein (vgl. GAP-DRG). Konnten anhand von weiteren Daten verwendet werden um Leistungen den Metaleistungen der NÖGKK zuzuordnen. Eine Aussage warum viele der Einträge nicht belegt sind kann allerdings nicht getroffen werden (~80% NULL Werte)

4.3 Fehlende Daten

Die Annotation der bestehenden Daten wurde mit größter Sorgfalt vorgenommen, dennoch konnten einige der enthaltenen Daten nicht ausreichend beschrieben werden. Diese werden im Folgenden zusammengefasst:

- Fachgruppen (siehe Tabelle `mat_fachgruppe`): die folgenden Codes für die Fachgruppen konnten bisher nicht geklärt werden: 76, 44, 45, 57, 79
- Leistungen (siehe Tabelle `lei_leistung`): wie beschrieben bleiben nach der Zuordnung der Leistungen zu den Metaleistungen immer noch ca. 2.833 Leistungspositionen ohne entsprechende Zuordnung über. Eine Nachbesserung der Daten in der Relation `lei_leistung` bzw. `mat_meta_leistung` wäre notwendig.
- Heilmittel (siehe Tabelle `lei_heilmittel`, Tabelle `mat_pharmazie`): im Datensatz sind weiterhin 255 Pharmazentralnummern enthalten welche noch nicht ausreichend annotiert werden konnten.

Abbildungsverzeichnis

Referenzen

1. Endel, F. *GAP-DRG Wiki: MBDS*. 2011 [cited 2018 2018-10-17]; Available from: http://gapdrg.endel.at/dokuwiki/doku.php/gapdrg:datenbank:datenmodell_inhalt:mbds.
2. Bundesanstalt Statistik Österreich. *Regionale Gliederung: Gemeinden*. 2018 [cited 2018 2018-10-17]; Available from: https://www.statistik.at/web_de/klassifikationen/regionale_gliederungen/gemeinden/index.html.
3. Bundesministerium für Gesundheit und Frauen, *Leistungsorientierte Krankenanstaltenfinanzierung Leistungskatalog BMGF 2018*. 2018.
4. Bundesministerium für Gesundheit und Frauen, *Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 10. Revision – BMGF-Version 2017.* 2016.
5. Deutsches Institut für Medizinische Dokumentation und Information. *Internationale Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-9)*. 1994; Available from: https://dimdi.de/static/de/klassi/icd-10-who/historie/icd-vorgaenger/icd-9/index.htm.
6. Bundesministerium für Gesundheit und Frauen. *Krankenanstalten in Österreich*. 2018 [cited 2018 2018-10-17]; Available from: https://www.sozialministerium.at/site/Gesundheit/Gesundheitssystem/Krankenanstalten/Krankenanstalten_Online-Verzeichnis/.
7. Bundesministerium für Arbeit, S., Gesundheit und Konsumentenschutz,. *Der Österreichische Strukturplan Gesundheit – ÖSG 2017*. 2017 [cited 2018 2018-10-17]; Available from: https://www.sozialministerium.at/site/Gesundheit/Gesundheitssystem/Gesundheitssystem_Qualitaetssicherung/Planung_und_spezielle_Versorgungsbereiche/Der_Oesterreichische_Strukturplan_Gesundheit_OeSG_2017.
8. Bundesministerium für Arbeit, S., Gesundheit und Konsumentenschutz,. *Zuordnung der politischen Bezirke und Gemeinden zu den Versorgungsregionen und Versorgungszonen des ÖSG*. 2017 [cited 2018-10-17; Available from: https://www.sozialministerium.at/cms/site/attachments/1/0/1/CH3967/CMS1136983382893/oesg_2017_-_regionale_gliederung.xlsx.
9. Bundesanstalt Statistik Österreich. *Regionale Gliederung: NUTS-Einheiten*. 2018 [cited 2018 2018-10-17]; Available from: https://www.statistik.at/web_de/klassifikationen/regionale_gliederungen/nuts_einheiten/index.html.