



Entwicklung einer ETL-Strecke zur Integration klinischer Daten in i2b2 mit Apache NiFi und Anbindung an eine PostgreSQL-Datenbank

15.04.2025 | MIRACUM-DIFUTURE

Erfan Matbouei

Core Unit Datenintegrationszentrum

Hinweis: Aus Gründen der Lesbarkeit wird in diesem Dokument das Generische Maskulinum verwendet, somit sind alle Formulierungen so zu verstehen, dass grundsätzlich immer alle Geschlechter eingeschlossen sind.

DIZ (Datenintegrationszentrum)



Warum i2b2 (Informatics for Integrating Biology & the Bedside)?

DIZ stellt Forschungsdaten bereit

Schnelle Info über verfügbare Daten

Prüfung der Eignung vor Anfrage



Was ist überhaupt i2b2?

- Suchmaschine für medizinische Daten
- Zeigt Patientenzahlen zu Merkmalen
- Keine Identifikation der Personen

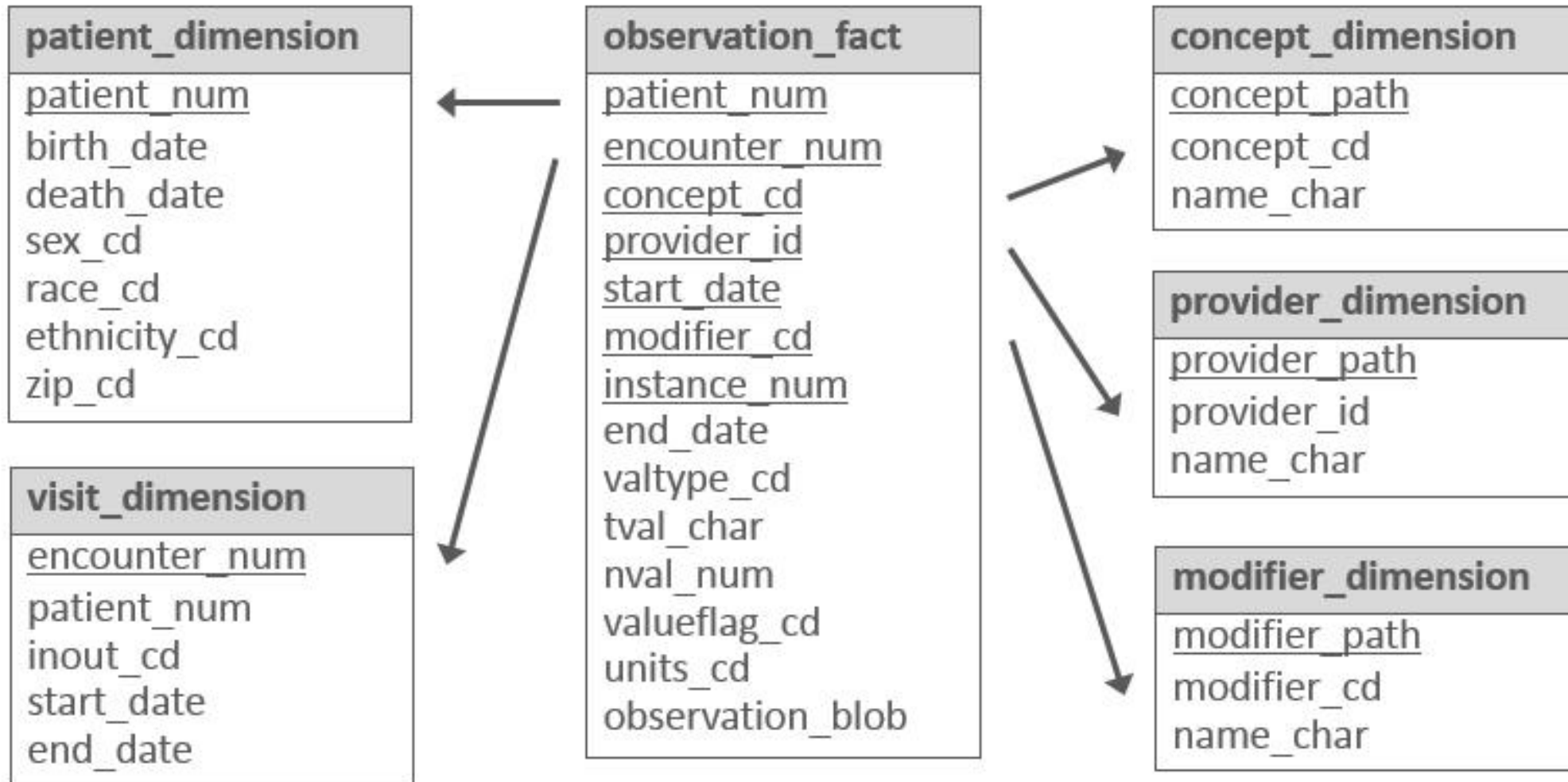


Wie ist i2b2 aufgebaut?

i2b2 basiert auf einem „**Faktenmodell**“, das grob in folgende Komponenten aufgeteilt ist:

| Komponente | Funktion |
|--------------------|--|
| Patient_dimension | Stammdaten von Patienten (anonymisiert) |
| Visit_dimension | Informationen zu Aufenthalten/Besuchen |
| Observation_fact | Einzelne medizinische Beobachtungen (z. B. Diagnosen, Laborwerte, Medikamente) |
| Concept_dimension | Enthält Metadaten zu klinischen Konzepten (z. B. Diagnosen, Medikamente, Laborwerte). |
| Provider_dimension | Optional: Behandelnde Personen oder Rollen |
| modifier_dimension | Diagnosen_arten, Wird verwendet, um zusätzliche Eigenschaften („Modifier“) zu einer Beobachtung (observation) zu speichern. Z. B. kann ein Labortest ein Modifier wie „Messmethode“ oder „Einheit“ haben. |
| i2b2 | Die i2b2 ist eine spezialisierte Erweiterung des i2b2-Datenmodells, die dazu dient, strukturierte Daten aus der klinischen Versorgung abzubilden. |

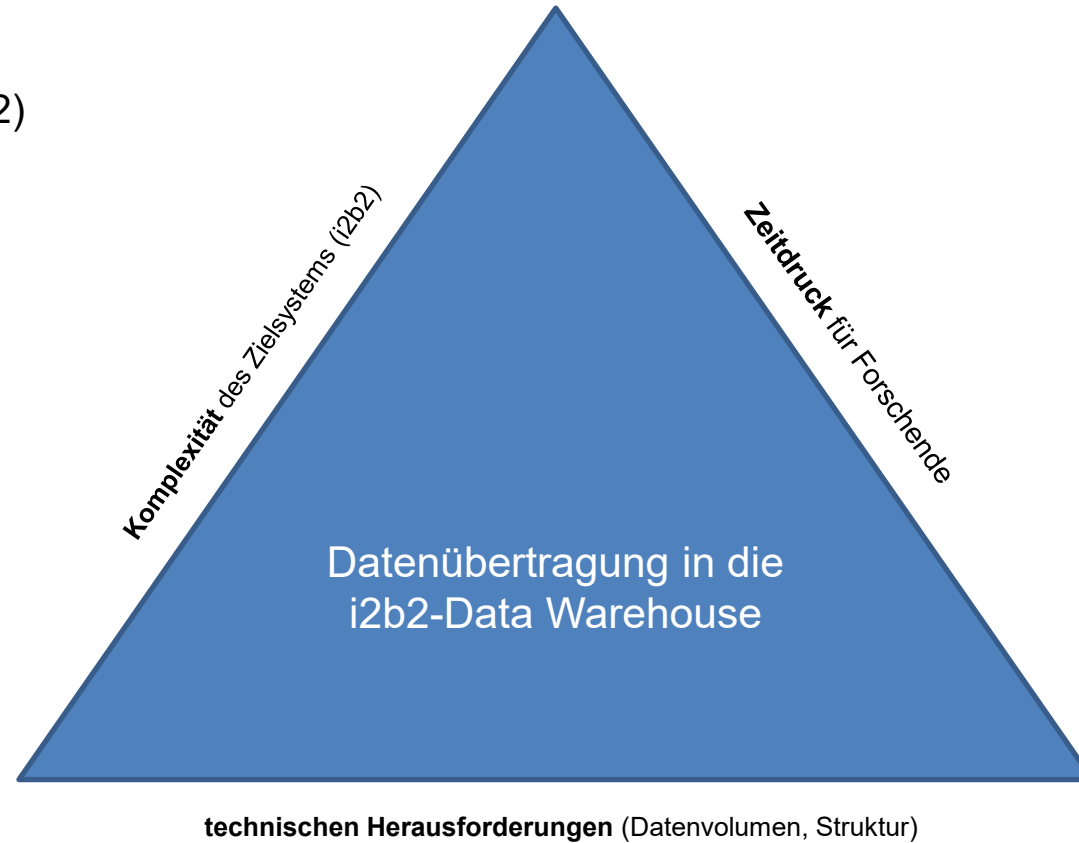
i2b2 Datenbank Star-Schema



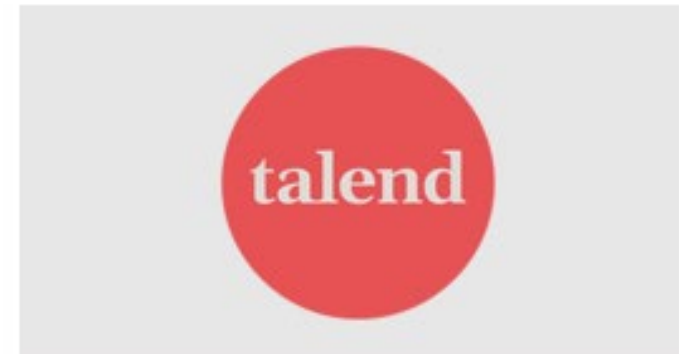
Was war das Problem bei der Datenübertragung?

Das Hauptproblem bei der Datenübertragung liegt in der Kombination aus:

1. **technischen Herausforderungen** (Datenvolumen, Struktur)
2. **Zeitdruck** für Forschende
3. **Komplexität** des Zielsystems (i2b2)



Gefundene Lösung: Einsatz eines ETL-Prozesses

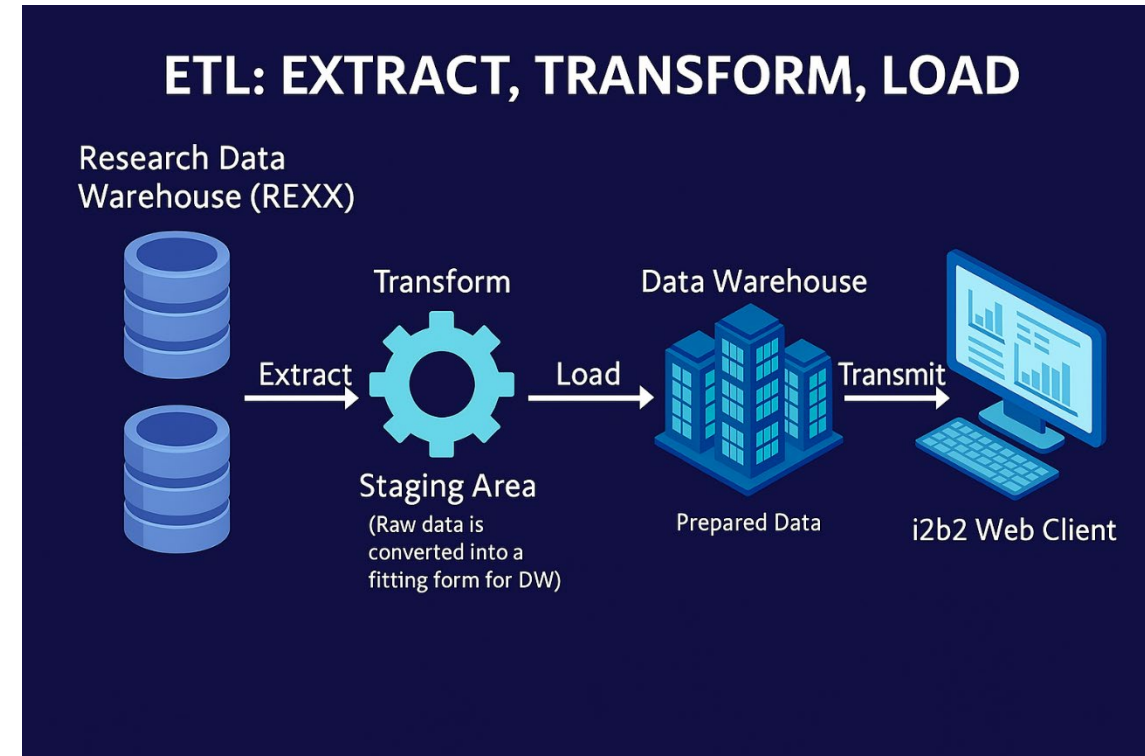


Was ist die Abkürzung ETL?

- **Extrahieren** von Daten aus verschiedenen externen Quellen
- **Transforming** in das erforderliche Geschäftsmodell
- **Loading** von Daten in das neue Data Warehouse

Warum funktioniert diese Lösung?

- **Automatisierung**
- **Standardisierung**
- **Zeitersparnis für Forschende**
- **Fehlerminimierung**



ETL Tool: Apache NiFi

Pricing: free

Official website: <https://nifi.apache.org/>

Useful resources: [documentation](#), [tutorials](#)

Pros:

- ✓ Perfect implementation of dataflow programming concept
- ✓ The opportunity to handle binary data
- ✓ Data provenance

Cons:

Simplistic UI



The screenshot shows the Apache NiFi website header with navigation links: Project, Documentation, Downloads, Community, Development, ASF Links, and Subprojects. Below the header is the Apache NiFi logo, which includes the text 'APACHE nifi' and a water drop icon. Underneath the logo is the tagline: 'An easy to use, powerful, and reliable system to process and distribute data.' To the right of the logo is a dataflow diagram. The diagram shows a flow from left to right. On the left, there are three 'To Tweets' processors, each with a status of 'Name health, analytic, unmatched' and 'Queued 0 (0 bytes)'. Arrows from these processors point to two analytics processors on the right. The top processor is 'Batch Analytics', which has a status of 'Queued 216,822 (506.84 MB)', 'In 53,564 (229.62 MB) → 6', 'Read/Write 0 bytes / 0 bytes', and 'Out 0 → 0 (0 bytes)'. The bottom processor is 'Streaming Analytics', which has a status of 'Queued 55,184 (236.96 MB)', 'In 0 (0 bytes) → 4', 'Read/Write 0 bytes / 0 bytes', and 'Out 0 → 0 (0 bytes)'. Both analytics processors have a status of 'No comments specified'.

Features

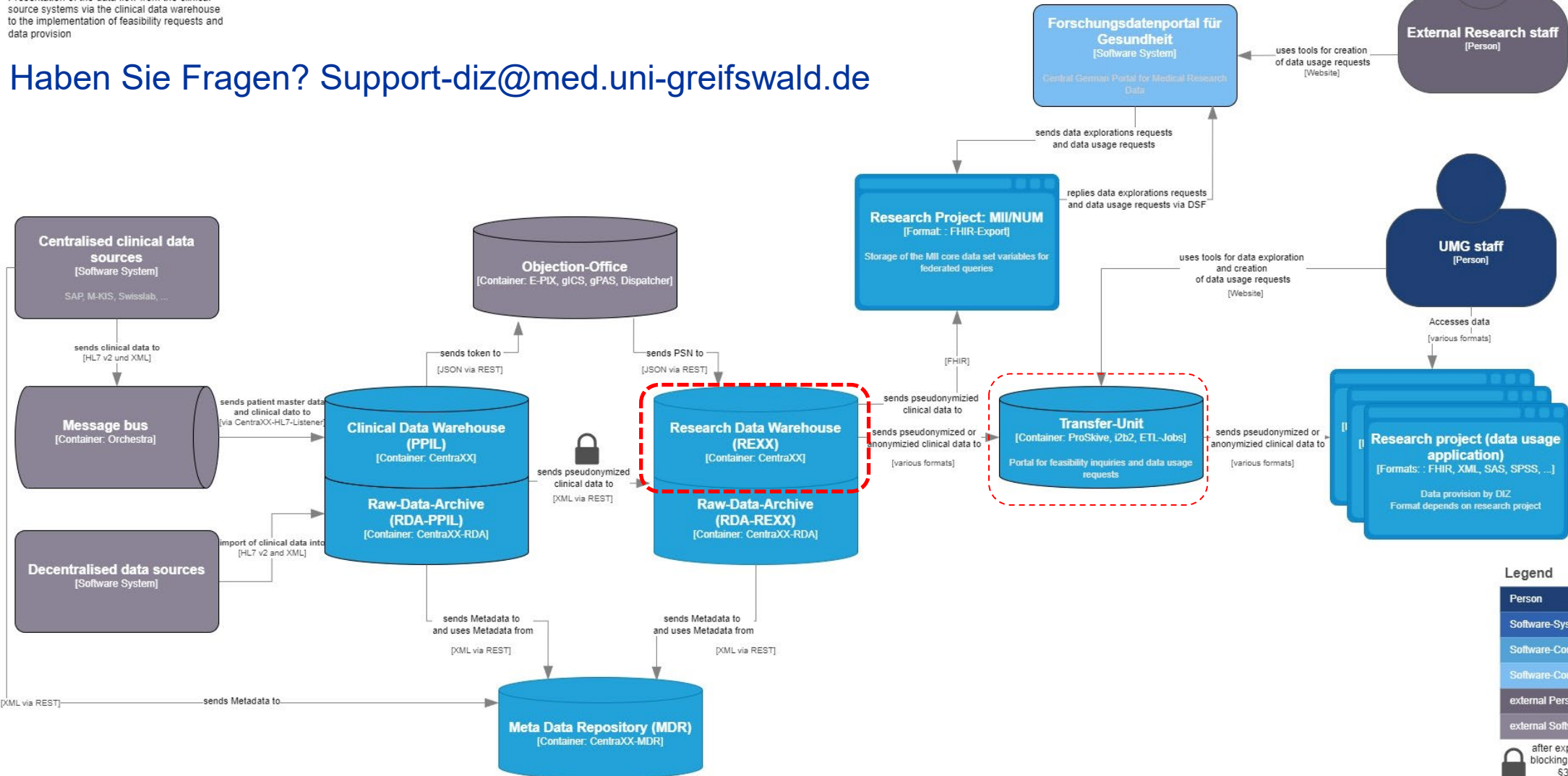
Apache NiFi supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic. Some of the high-level capabilities and objectives of Apache NiFi include:

DIZ IT-Architektur

Use case: Provision of clinical data for local and federated data use

Presentation of the data flow from the clinical source systems via the clinical data warehouse to the implementation of feasibility requests and data provision

Haben Sie Fragen? Support-diz@med.uni-greifswald.de



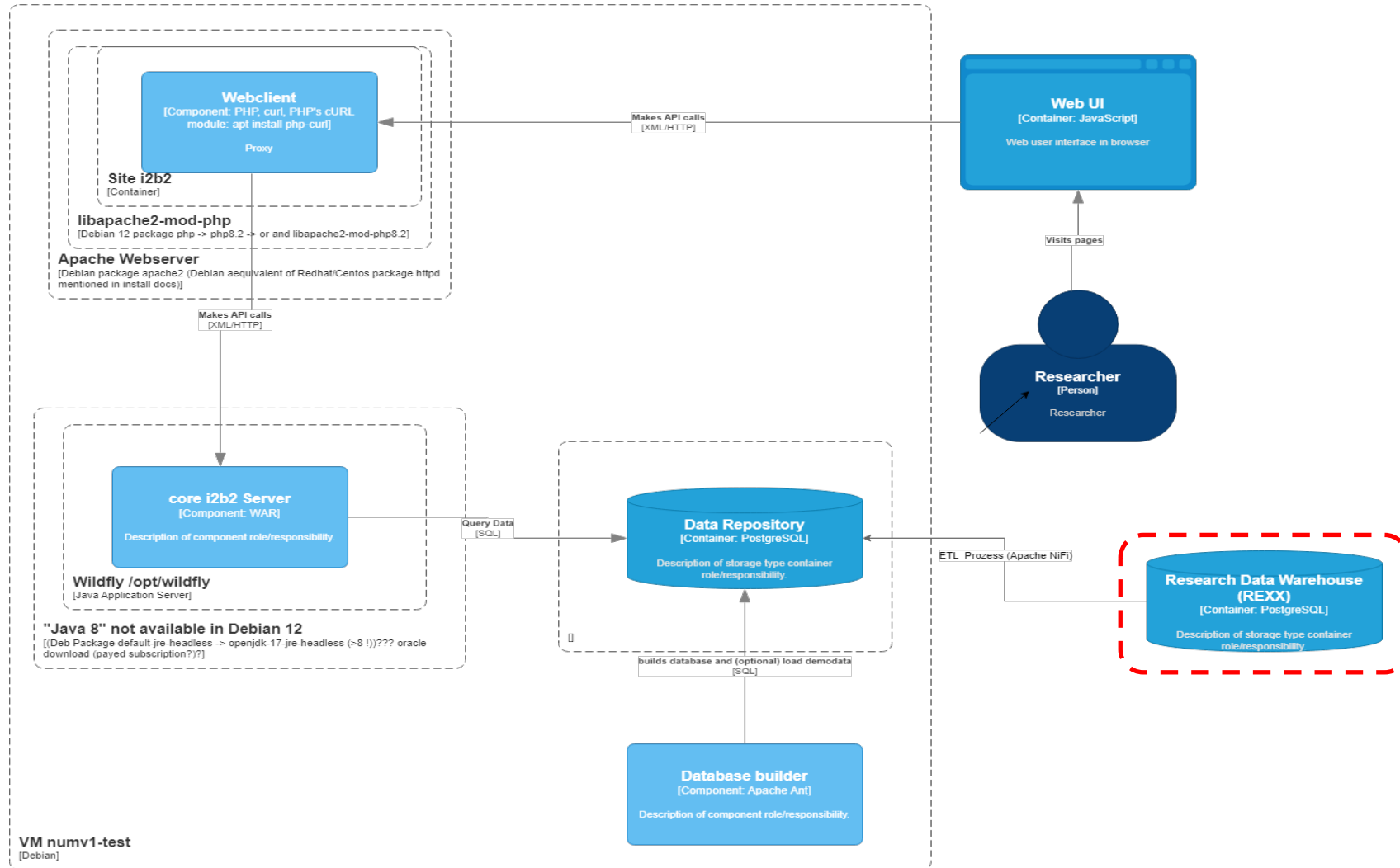
Legend

- Person
- Software-System DIC
- Software-Container DIC
- Software-Component DIC
- external Person
- external Software-System

after expiry of the 4-week blocking period LKHG MV §37 Section 5

i2b2 deployment diagram

[Deployment] Installation of i2b2 (for Redhat/Centos) on Debian VM of DIZ Greifswald
 siehe auch "Installation Guide" "Quick install" <https://www.i2b2.org/software/projects/hivecore/i2b2QuickInstall.pdf>

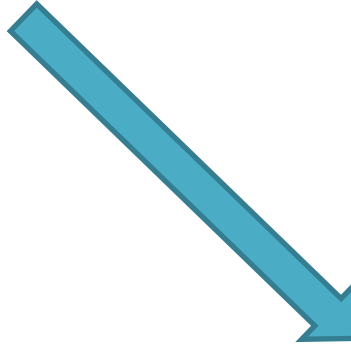


i2b2 Table and Table_access

Beispiel: Table_access

Data Output Messages Notifications

| | c_table_cd character varying (50) 🔒 | c_table_name character varying (50) 🔒 | c_protected_access character 🔒 | c_ontology_protection text 🔒 | c_hlevel integer 🔒 | c_fullname character varying (700) 🔒 | c_name character varying (2000) 🔒 |
|---|---|---|--|--|------------------------------|---|---|
| 1 | i2b2_DEMO | i2b2 | N | [null] | 1 | \i2b2\Demographics\ Demografische Daten | |
| 2 | i2b2 | i2b2 | N | [null] | 2 | \i2b2\Labor\SWISSLAB\ Laborergebnisse | |
| 3 | i2b2_ICD10 | i2b2 | N | [null] | 2 | \i2b2\diagnose\ICD-10-GM-DE_2024-GM-DE\ Diagnosen | |
| 4 | i2b2_OPS | i2b2 | N | [null] | 2 | \i2b2\Prozeduren\OPS_2024\ Prozeduren | |
| 5 | i2b2_M_KIS | i2b2 | N | [null] | 1 | \i2b2\M-KIS\ Dokumentationsseiten aus M-KIS (DOFI) | |



i2b2 Query & Analysis Tool

Terms ▾ Info

- + 📁 Demografische Daten
- + 📁 Diagnosen
- + 📁 Dokumentationsseiten aus M-KIS (DOFI)
- + 📁 Laborergebnisse
- + 📁 Prozeduren

Beispiel: i2b2

| c_hlevel integer | c_fullname character varying (700) | c_name character varying (2000) | c_synonym_cd character |
|---------------------|--|------------------------------------|---------------------------|
| 5 | \\i2b2\M-KIS\[0413] DOG.155233004\[0002] p2c2\[0001] CAVE\[0020] CAVE Medikamente\ | [0020] CAVE Medikamente | N |
| 4 | \\i2b2\M-KIS\[0413] DOG.155233004\[0001] p1c8\[0002] CAVE Medikamente\ | [0002] CAVE Medikamente | N |
| 5 | \\i2b2\M-KIS\[0413] DOG.155233004\[0001] p1c8\[0002] CAVE Medikamente\[0001] Medikamente\ | [0001] Medikamente | N |
| 6 | \\i2b2\M-KIS\[0413] DOG.155233004\[0001] p1c8\[0002] CAVE Medikamente\[0001] Medikamente\[0001] Medikamente\ | [0001] Medikamente | N |



Beispiel: i2b2

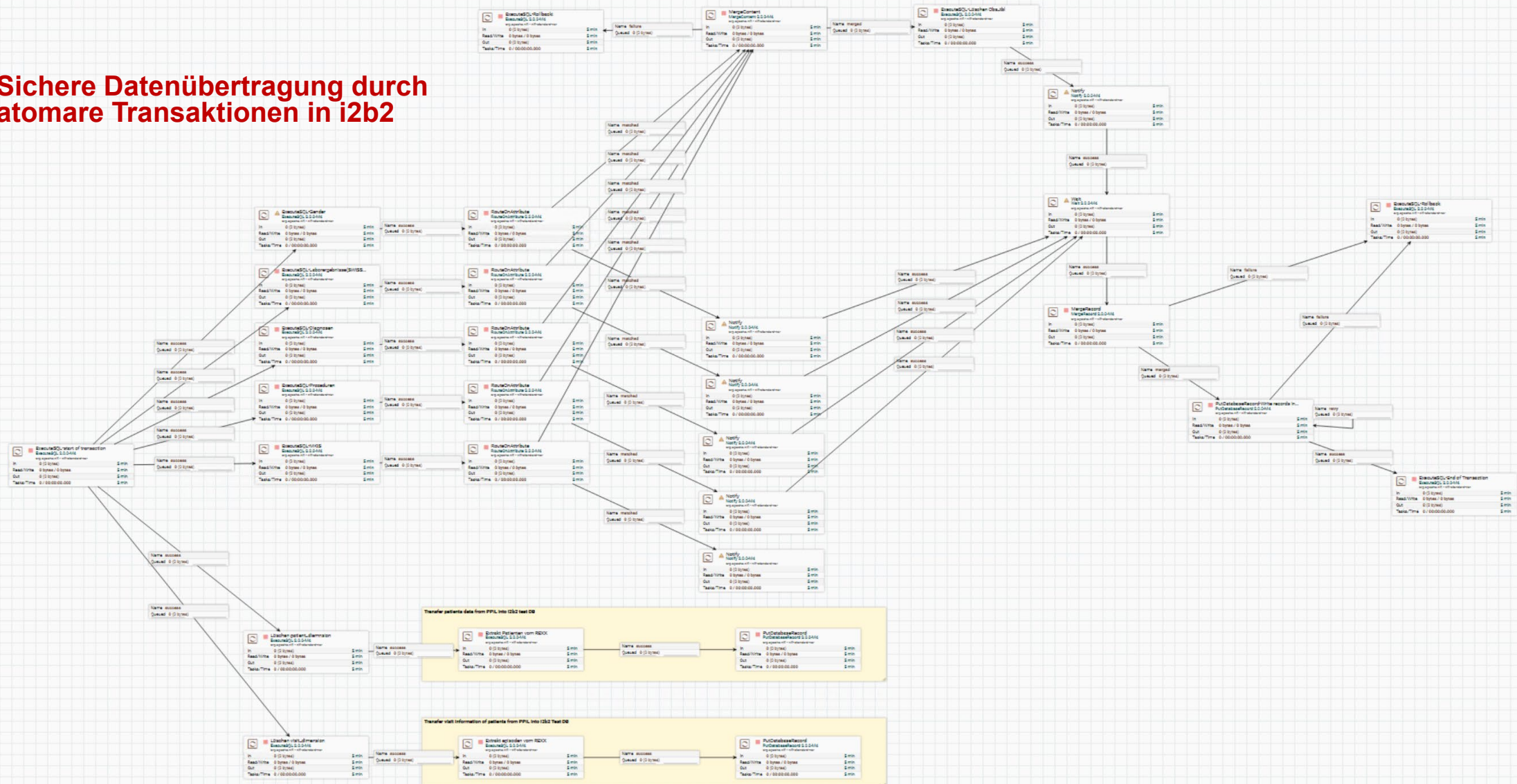
| | | | | | | |
|--|-----------------------|---|----|--------|--|--------------------------------------|
| <IS\[1168] Vitaldaten tabellarisch\[0001] p1c3\[0001] Vitaldaten tabellarisch\[0001] O2-Sättigung\ | [0001] O2-Sättigung | N | LA | [null] | DOF.1000101688.DOB.300000408.DOG.300000407(57b41bf8-4c58-44... | <?xml version="1.0" encoding="UTF-8" |
| <IS\[1168] Vitaldaten tabellarisch\[0001] p1c3\[0001] Vitaldaten tabellarisch\[0002] O2-Gabe\ | [0002] O2-Gabe | N | LA | [null] | DOF.1000101689.DOB.300000408.DOG.300000407(603113b7-064f-4e... | <?xml version="1.0" encoding="UTF-8" |
| <IS\[1168] Vitaldaten tabellarisch\[0001] p1c3\[0001] Vitaldaten tabellarisch\[0003] Gewicht\ | [0003] Gewicht | N | LA | [null] | DOF.1000101690.DOB.300000408.DOG.300000407(bc9f86c9-17a2-4b... | <?xml version="1.0" encoding="UTF-8" |
| <IS\[1168] Vitaldaten tabellarisch\[0001] p1c3\[0001] Vitaldaten tabellarisch\[0004] Blutzuckerwert\ | [0004] Blutzuckerwert | N | LA | [null] | DOF.1000101909.DOB.300000408.DOG.300000407(3c752bdf-e6ed-42... | <?xml version="1.0" encoding="UTF-8" |
| <IS\[1168] Vitaldaten tabellarisch\[0001] p1c3\[0001] Vitaldaten tabellarisch\[0005] Kostform\ | [0005] Kostform | N | LA | [null] | DOF.1000101691.DOB.300000408.DOG.300000407(2f2b2a8d-05f8-4b... | <?xml version="1.0" encoding="UTF-8" |



The screenshot shows a software interface with a hierarchical list of medical concepts on the left and a search dialog on the right. A blue arrow points from the table row for 'Kostform' to the corresponding entry in the list. The search dialog is titled 'Find Patients' and is currently set to search 'with' a concept. The concept selected is '[0005] Kostform'. Below this, there is a section for 'Choose value of Kostform' with a text input field containing a test ID. Below that, there are radio buttons for 'No value' and 'By value', with 'By value' selected. A dropdown menu is open, showing a list of values: 'Laktosefrei', 'Glutenfrei', 'Fortimel', 'Flüssigkost', and 'Angedickte Flüssigkeit'. At the bottom right of the dialog are 'Ok' and 'Cancel' buttons.

ETL Pipeline für i2b2 (Extract, Transform, Load)

**Sichere Datenübertragung durch
atomare Transaktionen in i2b2**



Start transaction!

| ExecuteSQL (Start of Transaction) | | |
|-------------------------------------|-------------------|-------|
| ExecuteSQL 2.2.0 | | |
| org.apache.nifi - nifi-standard-nar | | |
| In | 0 (0 bytes) | 5 min |
| Read/Write | 0 bytes / 0 bytes | 5 min |
| Out | 0 (0 bytes) | 5 min |
| Tasks/Time | 0 / 00:00:00.000 | 5 min |

1

Edit Processor | ExecuteSQL 2.2.0

Settings | Scheduling | **Properties** | Relationship

Required field + Verificati

| Property | Value |
|-------------------------------------|------------------------------|
| Database Connection Pooling Service | DBCPConnectionPool_i2b2_test |
| SQL Pre-Query | No value set |
| SQL select query | start transaction; |
| SQL Post-Query | No value set |

Click the component

2

3

Controller Service Details

DBCPConnectionPool 2.2.0

Settings | **Properties** | Comments

Required field

Verification

| Property | Value |
|-----------------------------|--|
| Database Connection URL | jdbc:postgresql://[redacted]/i2b2?stringtype=unspecified |
| Database Driver Class Name | |
| Database Driver Location(s) | |
| Kerberos User Service | |
| Database User | |
| Password | |

EL PARAM

Close

4

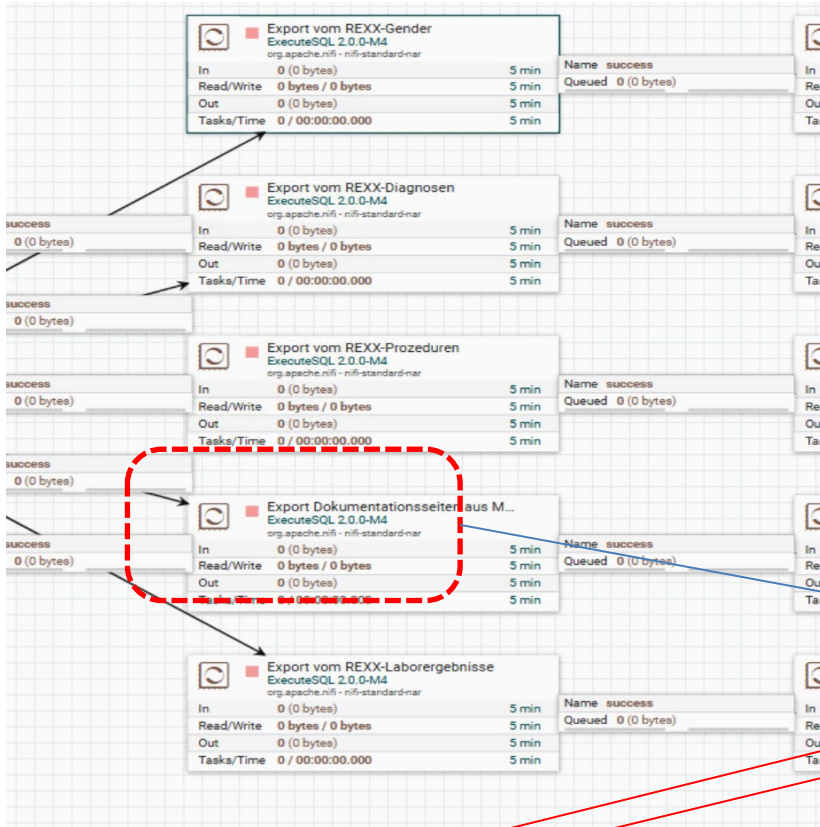
Controller Service Details

Settings | **Properties** | Verif

Required field

| Property | Value |
|-----------------------------|-------------------------------------|
| Database Connection URL | jdbc:postgresql://[redacted]/i2b... |
| Database Driver Class Name | org.postgresql.Driver |
| Database Driver Location(s) | C:\opt\ |
| Kerberos User Service | No value set |
| Database User | i2b2 |
| Password | Sensitive value set |

Extract und Transform(für Obs_fact Tabelle)



Edit Processor | ExecuteSQL 2.2.0

Settings Scheduling **Properties** Relationships Comments

Required field

+ Verification

Click the button above to verify this component.

| Property | Value |
|-------------------------------------|--------------------------------------|
| Database Connection Pooling Service | DBCConnectionPool_CXX_REXX |
| SQL Pre-Query | No value set |
| SQL select query | WITH MKIS AS (SELECT epi.episode... |
| SQL Post-Query | No value set |

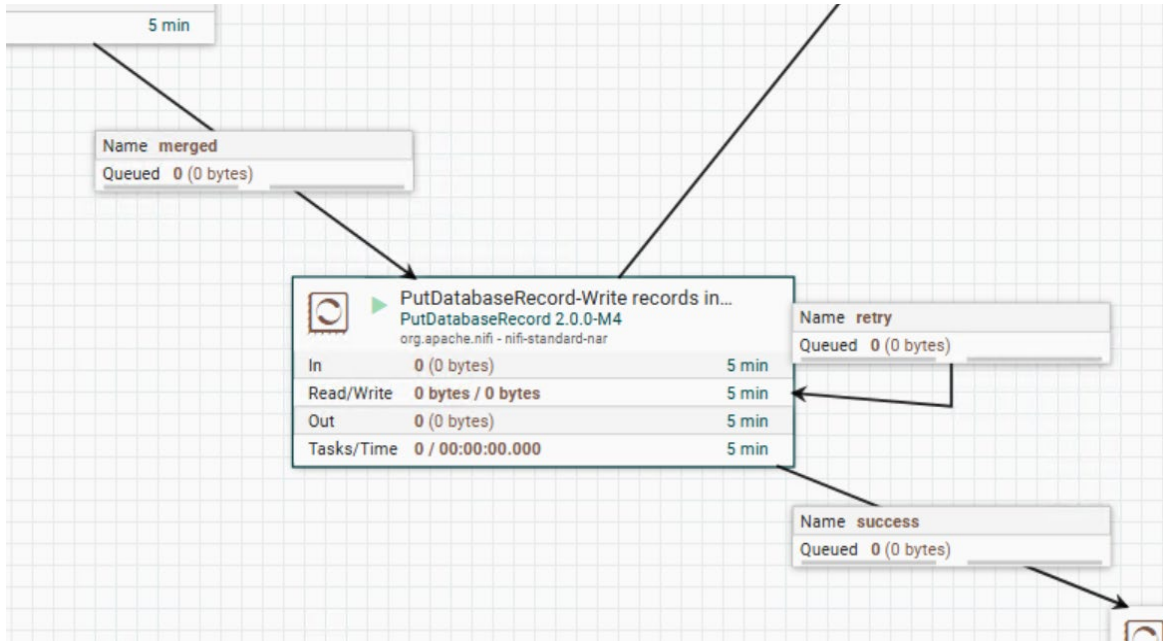
```

1 WITH MKIS AS (
2     SELECT
3         epi.episodeid AS ENCOUNTER_NUM,
4         pc.PATIENTID AS PATIENT_NUM,
5         lv.CODE AS concept_cd,
6         '@' AS provider_id,
7         lf.CREATIONDATE AS start_date,
8         '@' AS modifier_cd,
9         1 AS instance_num,
10        COALESCE(
11            CASE WHEN rv.STRINGVALUE IS NULL THEN NULL ELSE 'T' END,
12            CASE WHEN uade.VALUE IS NULL THEN NULL ELSE 'T' END,
13            CASE WHEN catde.VALUE IS NULL THEN NULL ELSE 'T' END,
14            CASE WHEN rv.NUMERICVALUE IS NULL THEN NULL ELSE 'N' END,
15            CASE WHEN rv.DATEVALUE IS NULL THEN NULL ELSE 'T' END,
16            CASE WHEN rv.DEVIANTVALUE IS NULL THEN NULL ELSE 'T' END
17        ) AS valtype_cd,
18        COALESCE(
19            rv.STRINGVALUE,
20            uade.VALUE,
21            catde.code,
22            CASE WHEN rv.NUMERICVALUE IS NULL THEN NULL ELSE 'E' END,
23            CASE WHEN rv.DATEVALUE IS NULL THEN NULL ELSE CAST(rv.DATEVALUE AS VARCHAR(25)) END,
24            CASE WHEN rv.DEVIANTVALUE IS NULL THEN NULL ELSE rv.DEVIANTVALUE END
25        ) AS tval_char,
26        COALESCE(
27            CAST(rv.NUMERICVALUE AS NUMERIC), NULL
28        ) AS nval_num,
29        lv.DTYPE AS valueflag_cd,
30        NOW() AS update_date,
31        NOW() AS import_date,
32        'M-KIS_CXX' AS sourcesystem_cd,
33        ROW_NUMBER() OVER (PARTITION BY epi.episodeid, pc.PATIENTID, lv.CODE ORDER BY lf.CREATIONDATE DESC) AS row_num
34    FROM ...

```

| encounter_num [PK] integer | patient_num [PK] integer | concept_cd [PK] character varying (10000) | provider_id [PK] character varying (255) | start_date [PK] timestamp without time zone | modifier_cd [PK] character | instance_num [PK] integer | valtype_cd character varying (255) | tval character |
|----------------------------|--------------------------|---|--|---|----------------------------|---------------------------|------------------------------------|----------------|
| 12 | 9 | DOF.1000101680.DOB.300000407.DOG.300000406(2a08ffb5-257a-4542-aae1-99ff4515a1f6) | @ | 2024-03-04 11:18:58.478 | @ | 1 | N | E |
| 12 | 9 | DOF.1000101676.DOB.300000407.DOG.300000406(d47d0380-8c86-4042-ad5a-9e1bdfb5278f) | @ | 2024-03-04 11:18:58.478 | @ | 1 | N | E |
| 12 | 9 | DOF.1000101684.DOB.300000407.DOG.300000406(4a1935fc-f66d-4dd5-9587-02bcd303802) | @ | 2024-03-04 11:18:58.478 | @ | 1 | N | E |
| 12 | 9 | DOF.1000101685.DOB.300000407.DOG.300000406(d9261ce2-637a-48e6-8085-4ab7ccdc0d0cb_v_9937b160-71b9-4013-9dc5-bd1d7c322..) | @ | 2024-03-04 11:18:58.478 | @ | 1 | T | A |
| 12 | 9 | DOF.1000101686.DOB.300000407.DOG.300000406(1b270d03-5c89-4485-ab9f-28a87d8fb480) | @ | 2024-03-04 11:18:58.478 | @ | 1 | N | E |
| 12 | 9 | DOF.1000101688.DOB.300000408.DOG.300000407(57b41bf8-4c58-44c9-88df-b8fcb5043438) | @ | 2024-03-04 11:30:58.361 | @ | 1 | N | E |
| 14 | 10 | DOF.1000101684.DOB.300000407.DOG.300000406(4a1935fc-f66d-4dd5-9587-02bcd303802) | @ | 2024-02-19 10:18:37.056 | @ | 1 | N | E |
| 14 | 10 | DOF.1000101676.DOB.300000407.DOG.300000406(d47d0380-8c86-4042-ad5a-9e1bdfb5278f) | @ | 2024-02-19 10:18:37.056 | @ | 1 | N | E |
| 14 | 10 | DOF.1000101678.DOB.300000407.DOG.300000406(10f9b164-6b56-4b64-9f2a-777bdfdec7bf) | @ | 2024-02-19 10:18:37.056 | @ | 1 | N | E |
| 14 | 10 | DOF.1000101680.DOB.300000407.DOG.300000406(2a08ffb5-257a-4542-aae1-99ff4515a1f6) | @ | 2024-02-19 10:18:37.056 | @ | 1 | N | E |

Load

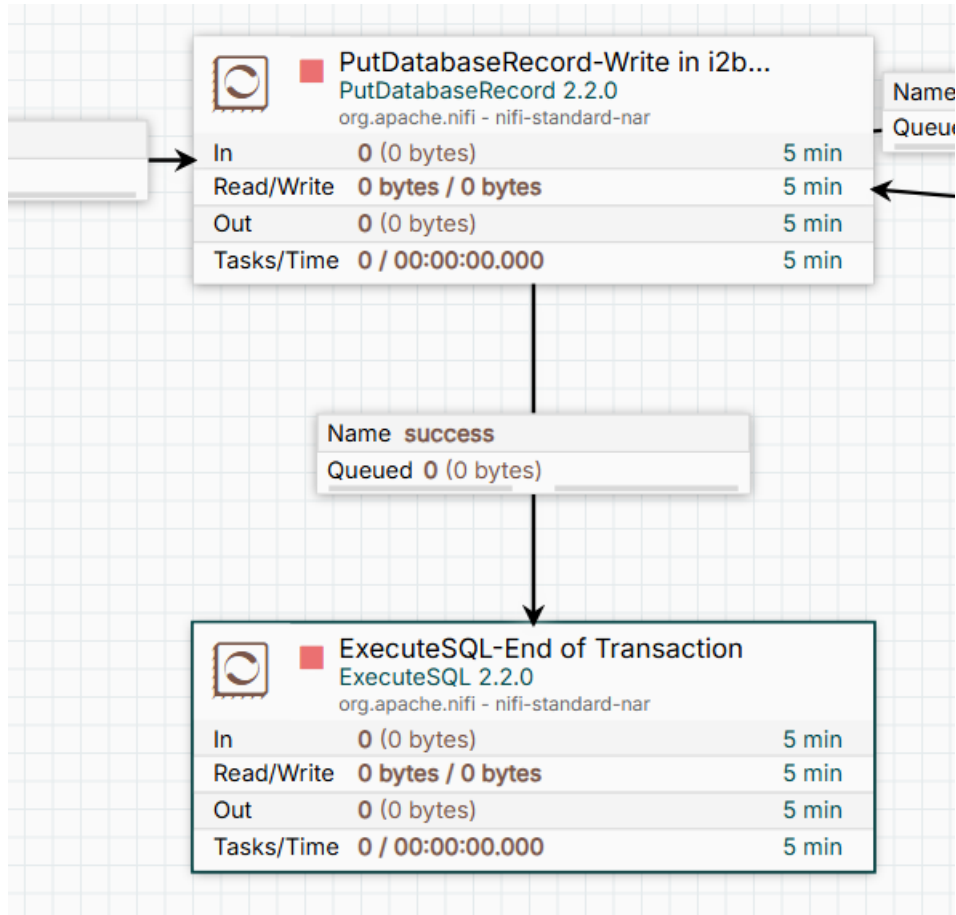


Required field

| Property | Value |
|-------------------------------------|--------------------------------------|
| Record Reader | AvroReader Warum Avro format? |
| Database Type | PostgreSQL |
| Statement Type | INSERT |
| Data Record Path | No value set |
| Database Connection Pooling Service | DBCPConnectionPool_i2b2_test |
| Catalog Name | No value set |
| Schema Name | No value set |
| Table Name | observation_fact |

**Tabelle als Zielstruktur
für die Datenbefüllung**

End of transaction



Edit Processor | ExecuteSQL 2.2.0

Settings

Scheduling

Properties

Required field

| Property | Value |
|-------------------------------------|-----------------------------|
| Database Connection Pooling Service | DBCConnectionPool_i2b2_test |
| SQL Pre-Query | No value set |
| SQL select query | commit; |

Anmeldung

i2b2 Login



i2b2 Host

Local and UMG Accounts

Username

scheuchm

Verwendung

Password

.....

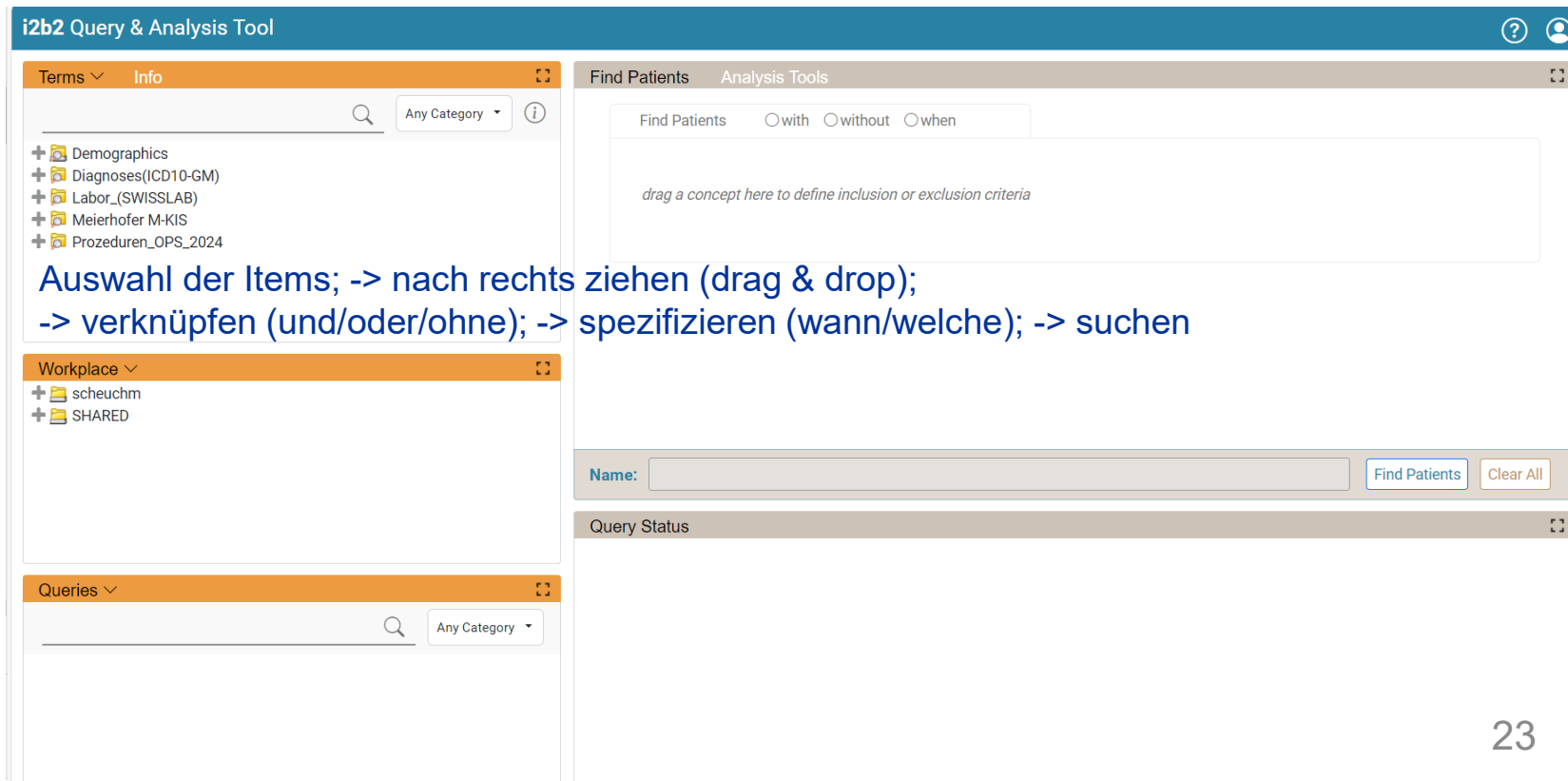
der Windows

Anmeldedaten

Login

Having trouble, please email [i2b2 Team](#)

i2b2- Weboberfläche

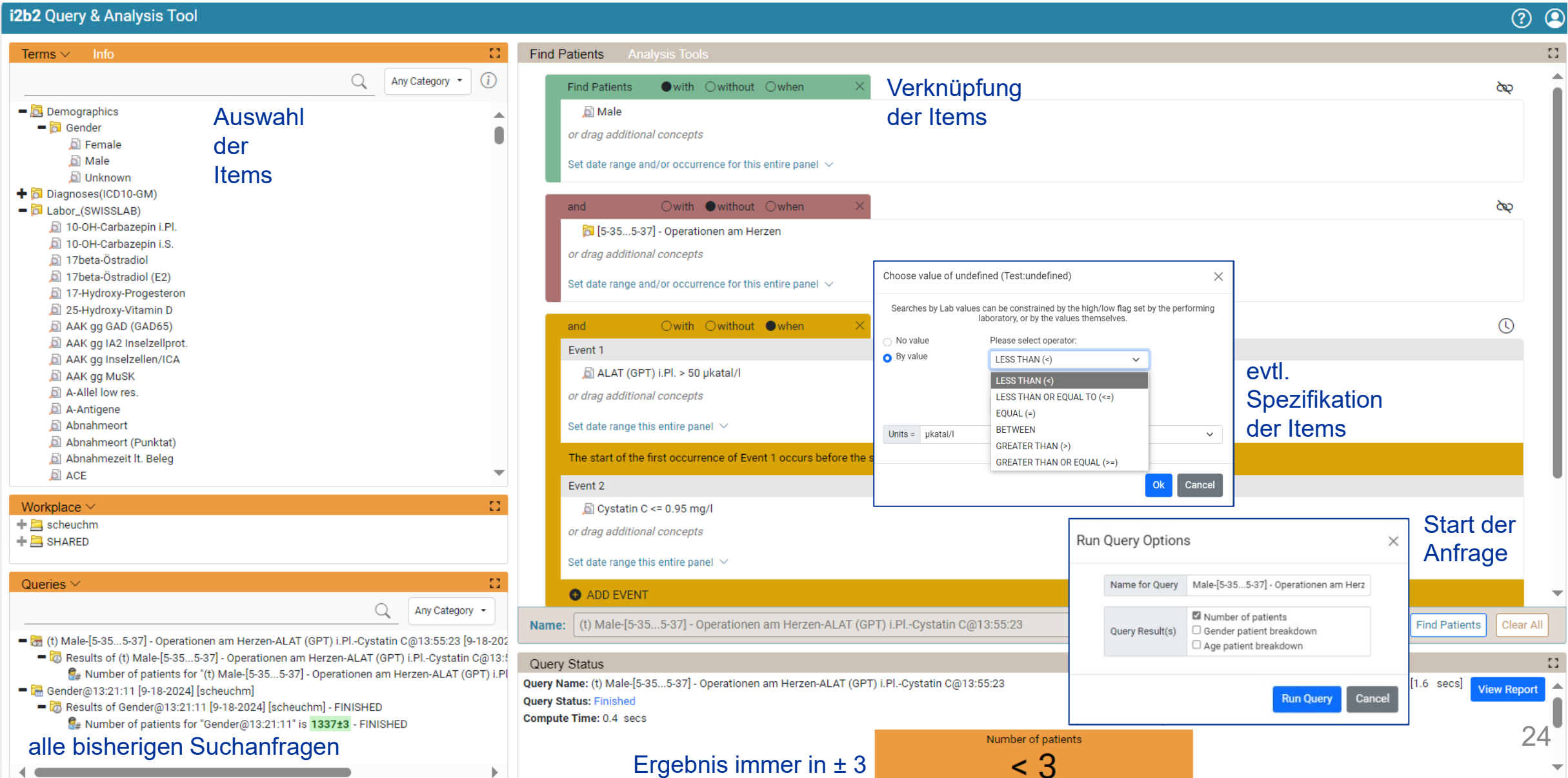


The screenshot shows the i2b2 Query & Analysis Tool interface. It features a sidebar with a tree view of data sources under 'Terms' and 'Workplace'. The 'Terms' section includes categories like Demographics, Diagnoses(ICD10-GM), Labor_(SWISSLAB), Meierhofer M-KIS, and Prozeduren_OPS_2024. The 'Workplace' section shows 'scheuchm' and 'SHARED'. The main area is divided into 'Find Patients' and 'Analysis Tools' tabs. The 'Find Patients' tab has a search bar and a 'Find Patients' button. Below it, there is a 'Name:' field and a 'Clear All' button. The 'Query Status' section is currently empty.

Auswahl der Items; -> nach rechts ziehen (drag & drop);
-> verknüpfen (und/oder/ohne); -> spezifizieren (wann/welche); -> suchen

Beispielsuche

Männliche Patienten, ohne OP am Herzen, mit einem erhöhten ALAT, wenn gleichzeitig Cystatin C im Normalbereich



Terms Info

- Demographics
 - Gender
 - Female
 - Male
 - Unknown
- Diagnoses (ICD10-GM)
- Labor_ (SWISSLAB)
 - 10-OH-Carbazepin i.PI.
 - 10-OH-Carbazepin i.S.
 - 17beta-Östradiol
 - 17beta-Östradiol (E2)
 - 17-Hydroxy-Progesteron
 - 25-Hydroxy-Vitamin D
 - AAK gg GAD (GAD65)
 - AAK gg IA2 Inselzellprot.
 - AAK gg Inselzellen/ICA
 - AAK gg MuSK
 - A-Allel low res.
 - A-Antigene
 - Abnahmeort
 - Abnahmeort (Punktat)
 - Abnahmezeit lt. Beleg
 - ACE

Workplace

- scheuchm
- SHARED

Queries

- (t) Male-[5-35...5-37] - Operationen am Herzen-ALAT (GPT) i.PI.-Cystatin C@13:55:23 [9-18-2024]
- Results of (t) Male-[5-35...5-37] - Operationen am Herzen-ALAT (GPT) i.PI.-Cystatin C@13:55:23 [9-18-2024]
- Gender@13:21:11 [9-18-2024] [scheuchm]
- Results of Gender@13:21:11 [9-18-2024] [scheuchm] - FINISHED
- Number of patients for "Gender@13:21:11" is **1337±3** - FINISHED

Auswahl der Items

Find Patients Analysis Tools

Find Patients with without when

Male

or drag additional concepts

Set date range and/or occurrence for this entire panel

and with without when

[5-35...5-37] - Operationen am Herzen

or drag additional concepts

Set date range and/or occurrence for this entire panel

and with without when

Event 1

ALAT (GPT) i.PI. > 50 µkatal/l

or drag additional concepts

Set date range this entire panel

The start of the first occurrence of Event 1 occurs before the s

Event 2

Cystatin C <= 0.95 mg/l

or drag additional concepts

Set date range this entire panel

ADD EVENT

Verknüpfung der Items

Choose value of undefined (Test:undefined)

Searches by Lab values can be constrained by the high/low flag set by the performing laboratory, or by the values themselves.

No value

By value

Please select operator:

- LESS THAN (<)
- LESS THAN OR EQUAL TO (<=)
- EQUAL (=)
- BETWEEN
- GREATER THAN (>)
- GREATER THAN OR EQUAL (>=)

Units = µkatal/l

Ok Cancel

evtl. Spezifikation der Items

Run Query Options

Name for Query Male-[5-35...5-37] - Operationen am Herz

Query Result(s)

- Number of patients
- Gender patient breakdown
- Age patient breakdown

Run Query Cancel

Start der Anfrage

Name: (t) Male-[5-35...5-37] - Operationen am Herzen-ALAT (GPT) i.PI.-Cystatin C@13:55:23

Query Status

Query Name: (t) Male-[5-35...5-37] - Operationen am Herzen-ALAT (GPT) i.PI.-Cystatin C@13:55:23

Query Status: Finished

Compute Time: 0.4 secs

Number of patients

Ergebnis immer in ± 3

< 3

Find Patients Clear All

1.6 secs View Report

24

alle bisherigen Suchanfragen

Zusammenfassung – Kernaussagen des Vortrags

- **i2b2** ist ein leistungsfähiges Recherchetool für medizinische Daten zur Unterstützung der Forschung.
- **Forschende** können vorab prüfen, ob relevante Daten verfügbar sind – ohne aufwändige Anträge.
- Die **Datenintegration** in i2b2 war technisch herausfordernd (Datenvolumen, Komplexität, Zeitdruck).
- Die Lösung war ein automatisierter **ETL-Prozess mit Apache NiFi**, der Daten standardisiert, sicher und effizient überträgt.
- Die **i2b2-Weboberfläche** ermöglicht einfache und schnelle Abfragen mit Drag & Drop – in wenigen Sekunden zum Ergebnis.
- Durch die i2b2-Plattform wird der Zugang zu Forschungsdaten transparenter, schneller und niederschwelliger.
- verstehen die Komponenten und die Funktionsweise der ETL-Strecke zur Integration klinischer Daten in i2b2

Vielen Dank für Ihre Aufmerksamkeit!



Erfan Matbouei
Universitätsmedizin Greifswald (K.d.ö.R.)
Core Unit Datenintegrationszentrum

<https://www.medizin.uni-greifswald.de/diz/>

