Pseudonymisierung und Anonymisierung von klinischen Daten in den Datenintegrationszentren der MII

Matthias Löbe
Institut für Medizinische Informatik, Statistik und Epidemiologie (IMISE),
Universität Leipzig
matthias.loebe@imise.uni-leipzig.de

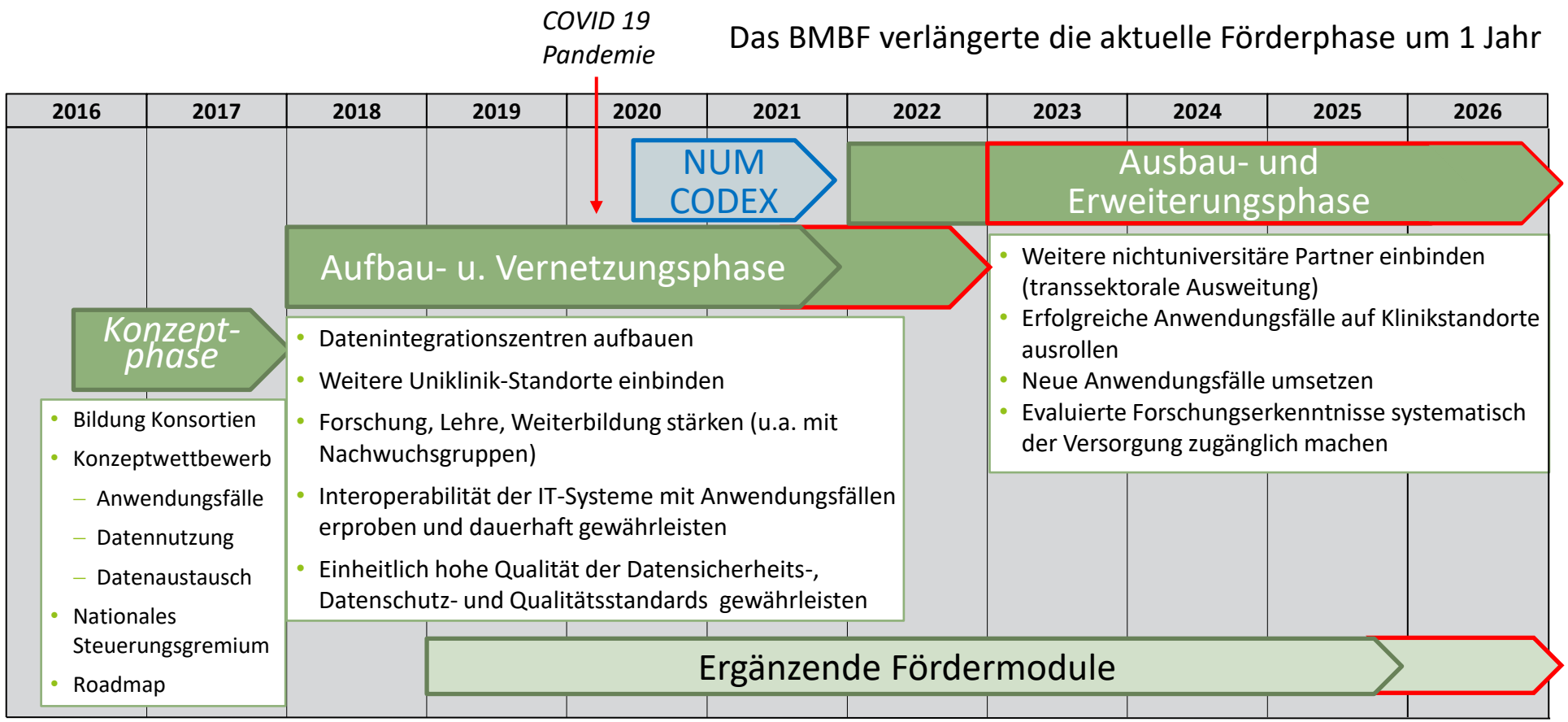
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Hintergrund Medizininformatik-Initiative (MII)

- Förderkonzept 2015 vom Bundesministerium für Bildung und Forschung initiiert
- Förderung in 3 Phasen von 2016-2026
- Fördersumme rund 400 Millionen Euro
- 37 universitätsmedizinische Standorte und weitere Partner bundesweit

https://www.medizininformatik-initiative.de/

Die übergreifende Zusammenarbeit in der aktuell laufenden Aufbau- und Vernetzungsphase orientiert sich an der Roadmap der MII



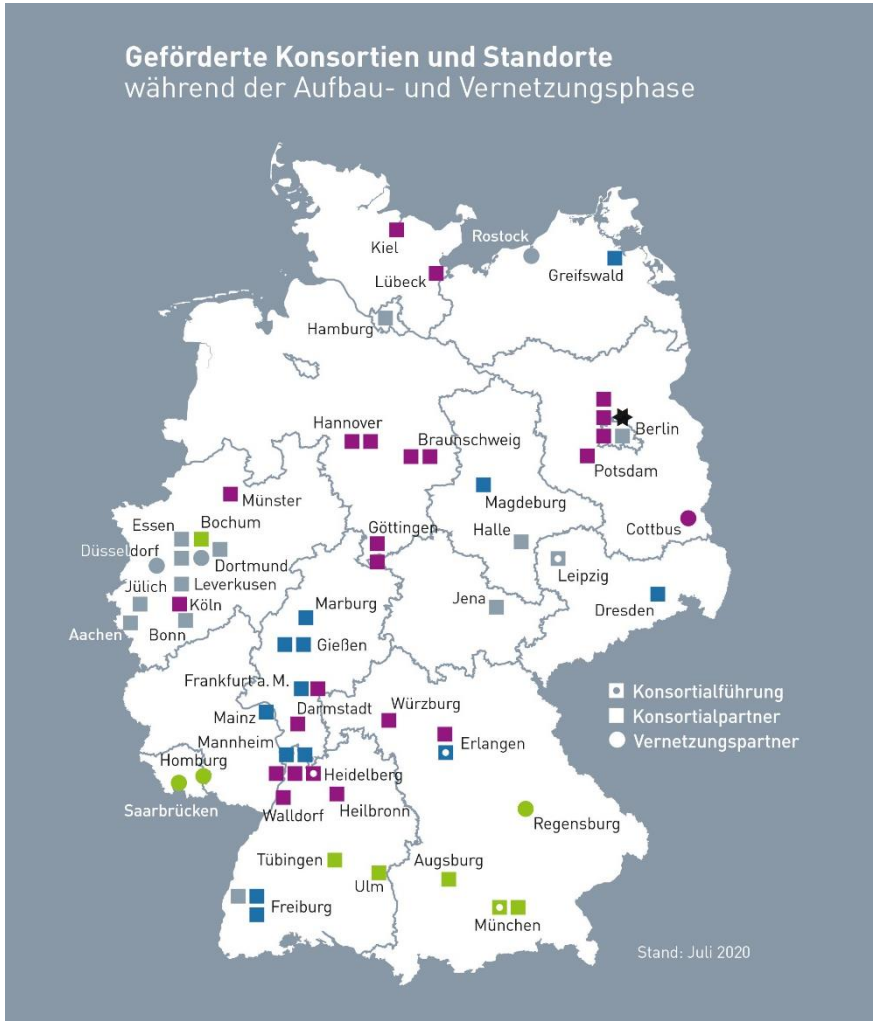
→ **Mitten in der Aufbau- und Vernetzungsphase wurden Ressourcen der MII kurzfristig in das Projekt CODEX umgeleitet**

Konsortien und Datenintegrationszentren

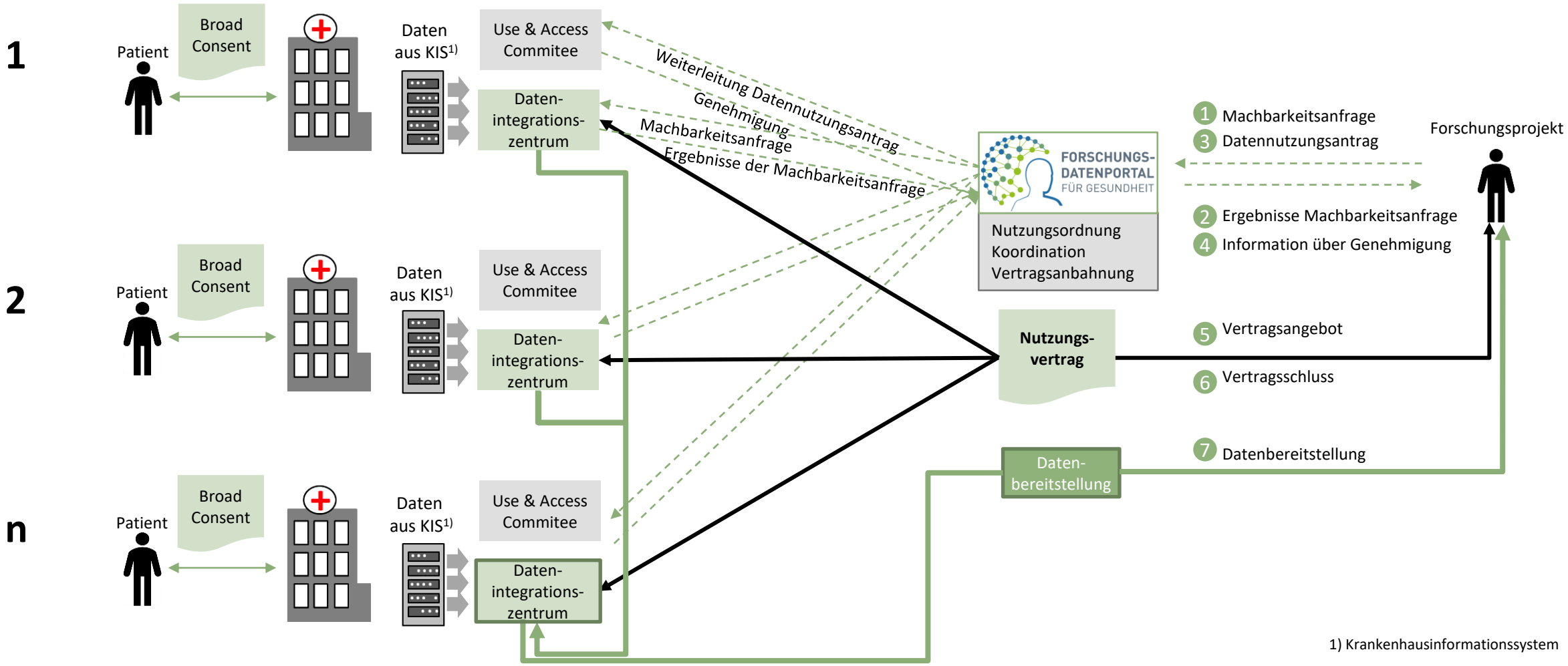
- 37 universitätsmedizinische Standorte sind der MII in der Aufbau- und Vernetzungsphase angeschlossen

- **DIFUTURE** 7 Standorte
- **HiGHmed** 10 Standorte
- **MIRACUM** 10 Standorte
- **SMITH** 10 Standorte

- **Koordinationsstelle:**
  TMF, MFT, VUD



Geplanter Data-Sharing-Prozess in der MII



Projectathons sind ein wichtiges Instrument zur praxisnahen Überprüfung der Abläufe und Strukturen des Data Sharing

Projectathons der MII sind konsortienübergreifend

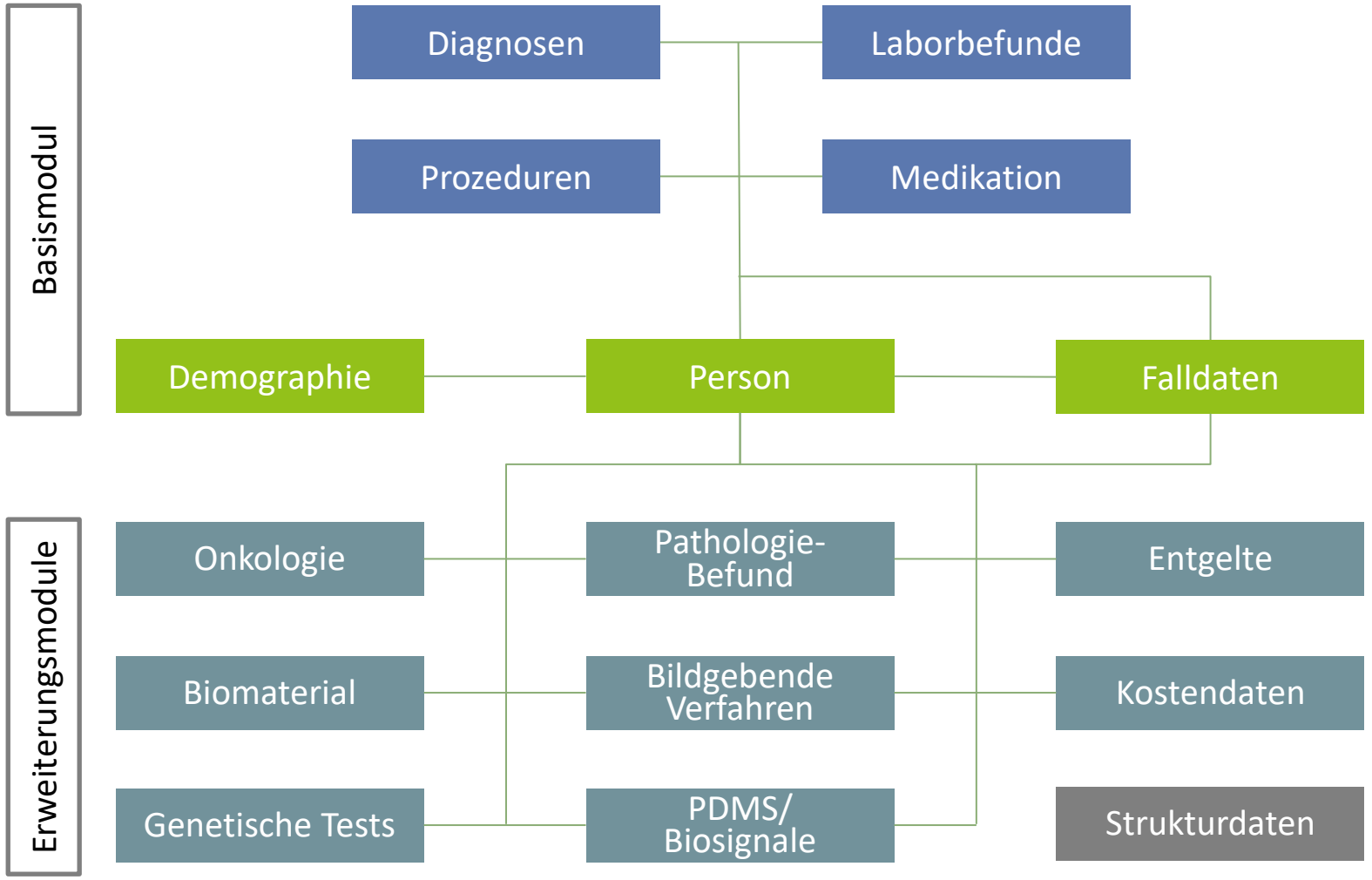
- Prozessen und Strukturen gemeinsam erproben
- Schwachstellen und Fehler gemeinsam identifizieren
- Lösungen teilen oder miteinander entwickeln
- Data-Sharing-Infrastruktur gemeinsam verbessern und weiterentwickeln

1. Projectathon	Herbst 2019	▸ Erprobung der Module Labor, Medikation und Person
2. Projectathon	Sommer 2020	▸ Erprobung der Module Fall, Diagnose, Prozedur
3. Projectathon	Herbst 2020	▸ Erprobung Machbarkeitsanfrage Stufe 1 /Durchführung erste Datenabfragen mit FHIR Search ▸ Überführung von Daten aus dem FHIR-Format in auswertbare Formate
4. Projectathon	Frühjahr 2021	▸ Erprobung Machbarkeitsanfrage Stufe 1 /Durchführung erste Datenabfragen mit FHIR Search ▸ Einbindung des Forschungsdatenportals in den Prozessablauf
5. Projectathon	Sommer 2021	▸ Erster Durchlauf einer externen Machbarkeitsuntersuchung ▸ Erste Testabfrage für das Schaufenster des Deutschen Forschungsdatenportals für Gesundheit.
6. Projectathon	Herbst 2021	▸ Erprobung des Antrags-, Vertrags- und Datenbereitstellungsprozesses (manuell)
7. Projectathon	Herbst 2022	▸ Erprobung des Antrags-, Vertrags- und Datenbereitstellungsprozesses mit Forschungsdatenportal und Data Sharing Framework

Datenkörper: Daten aus dem Electronic Health Record

„Kerndatensatz“

- *Basismodule*: Daten, die jedes DIZ bereitstellen muss
- *Erweiterungsmodule*: Daten, die harmonisiert bereitgestellt werden, falls vorhanden
- *Governance*: Fest-legungen zu Aufnahme, Pflege und Versionierung von Inhalten



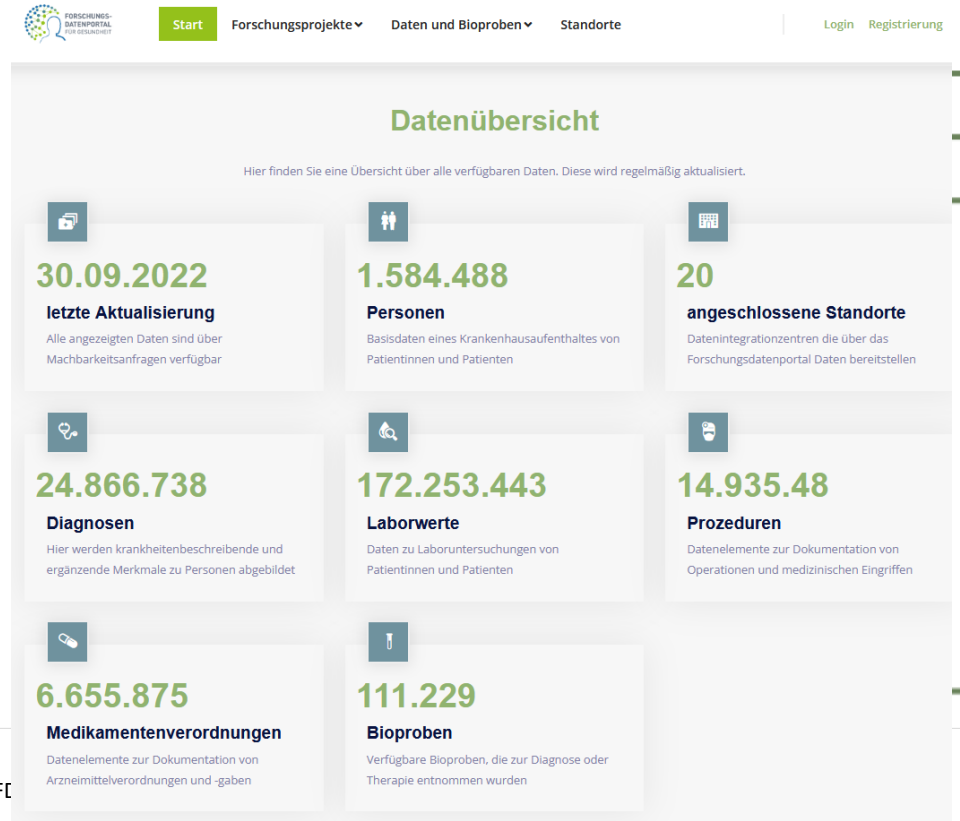
Gemeinsames Datenformat: FHIR

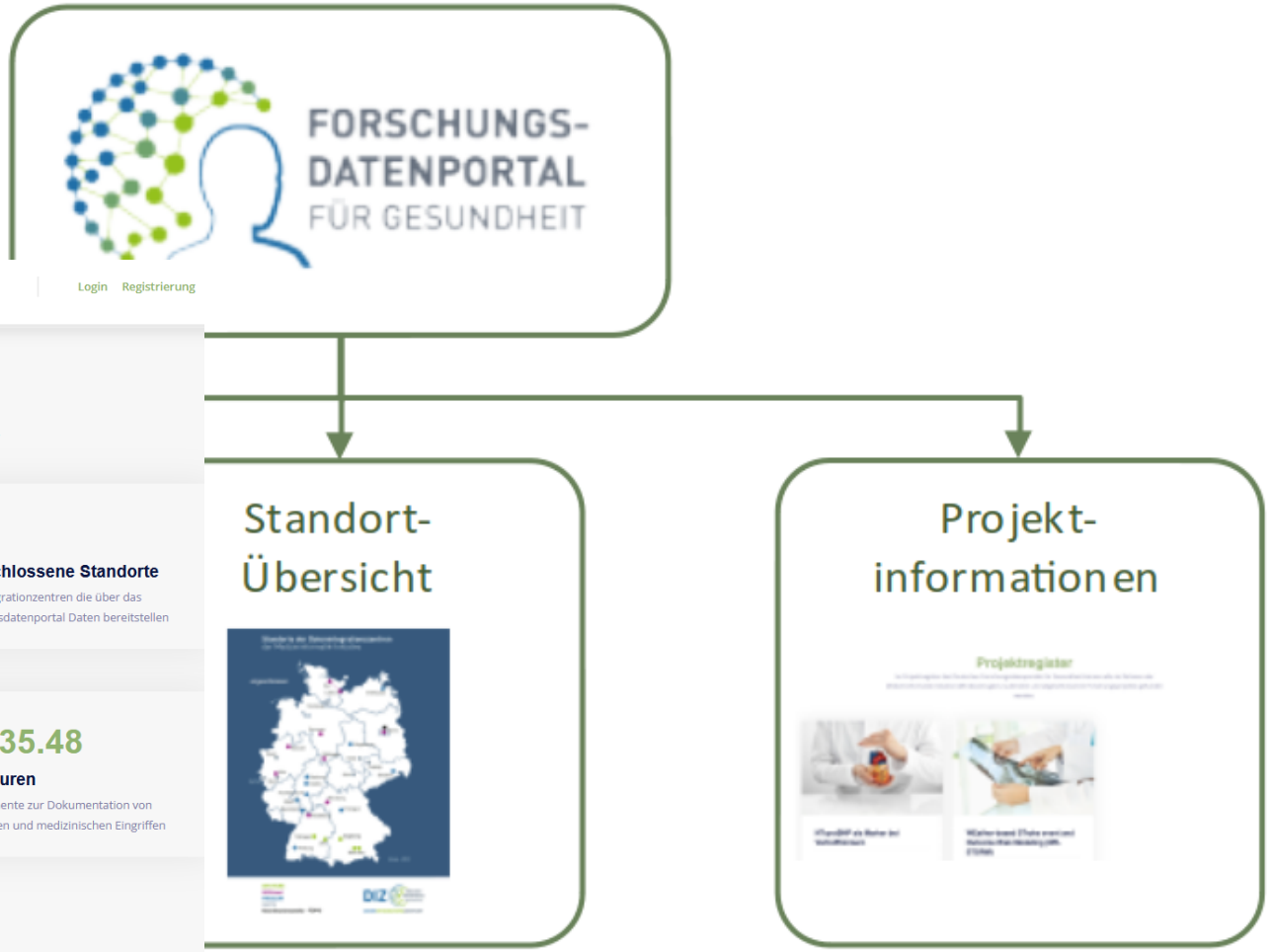
- MII setzt auf internationale Standards
  - 4 Konsortien = 4 verschiedene Ansätze
  - HL7 CDA, IHE, OHDSI OMOP, Harvard i2b2, openEHR Archetypes, CDISC ODM
- Einigung auf HL7 FHIR als zentrales Austauschformat
  - Unterstützung von allen Konsortien
  - Zu Projektbeginn relativ neu und ungetestet
- HL7 FHIR ist State of the Art
  - Sehr komplexe, hierarchische Struktur
  - Erlaubt sehr granulare Modellierung der vorhandenen Daten
  - Nutzt internationale Vokabulare und Standardterminologien wie LOINC und SNOMED CT
- HL7 FHIR ist nicht CSV!
  - Rückgabe als XML oder JSON


```
},
"issued": "2013-04-04T14:34:00+01:00",
"performer": [
  {
    "reference": "Practitioner/f202",
    "display": "Luigi Maas"
  }
],
"valueQuantity": {
  "value": 122,
  "unit": "umol/L",
  "system": "http://snomed.info/sct",
  "code": "258814008"
},
"interpretation": [
  {
    "coding": [
      {
        "system": "http://snomed.info/sct",
        "code": "166717003",
        "display": "Serum creatinine raised"
      },
      {
        "system": "http://terminology.hl7.org/CodeSystem/v3-ObservationInterpretation",
        "code": "H"
      }
    ]
  }
],
"referenceRange": [
  {
    "low": {
      "value": 64
    },
    "high": {
      "value": 104
    },
    "type": {
      "coding": [
        {
          "system": "http://terminology.hl7.org/CodeSystem/referencerange-meaning",
          "code": "normal",
          "display": "Normal Range"
        }
      ]
```

Forschungsdatenportal Gesundheit – öffentliche Sicht

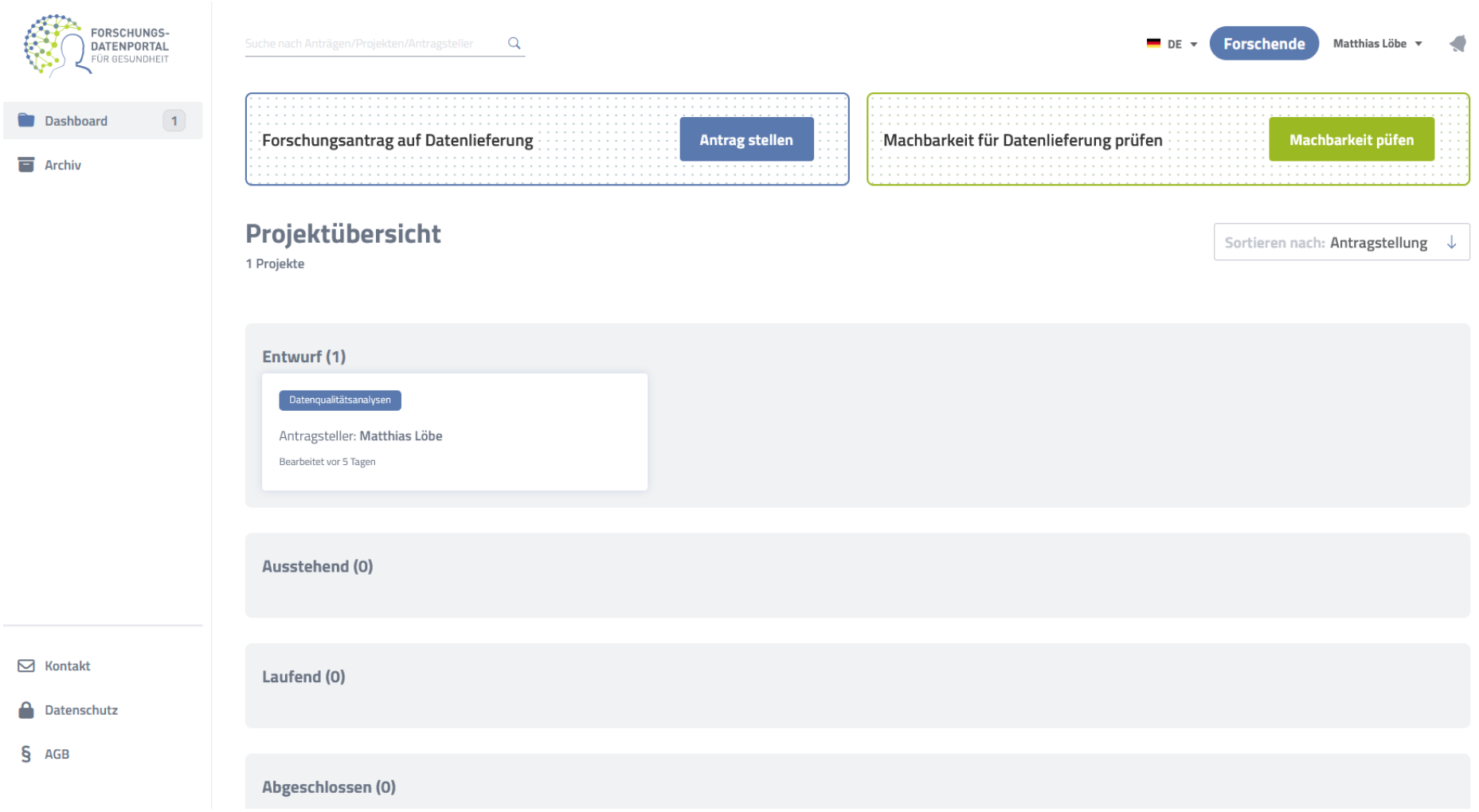
Not great, not terrible.



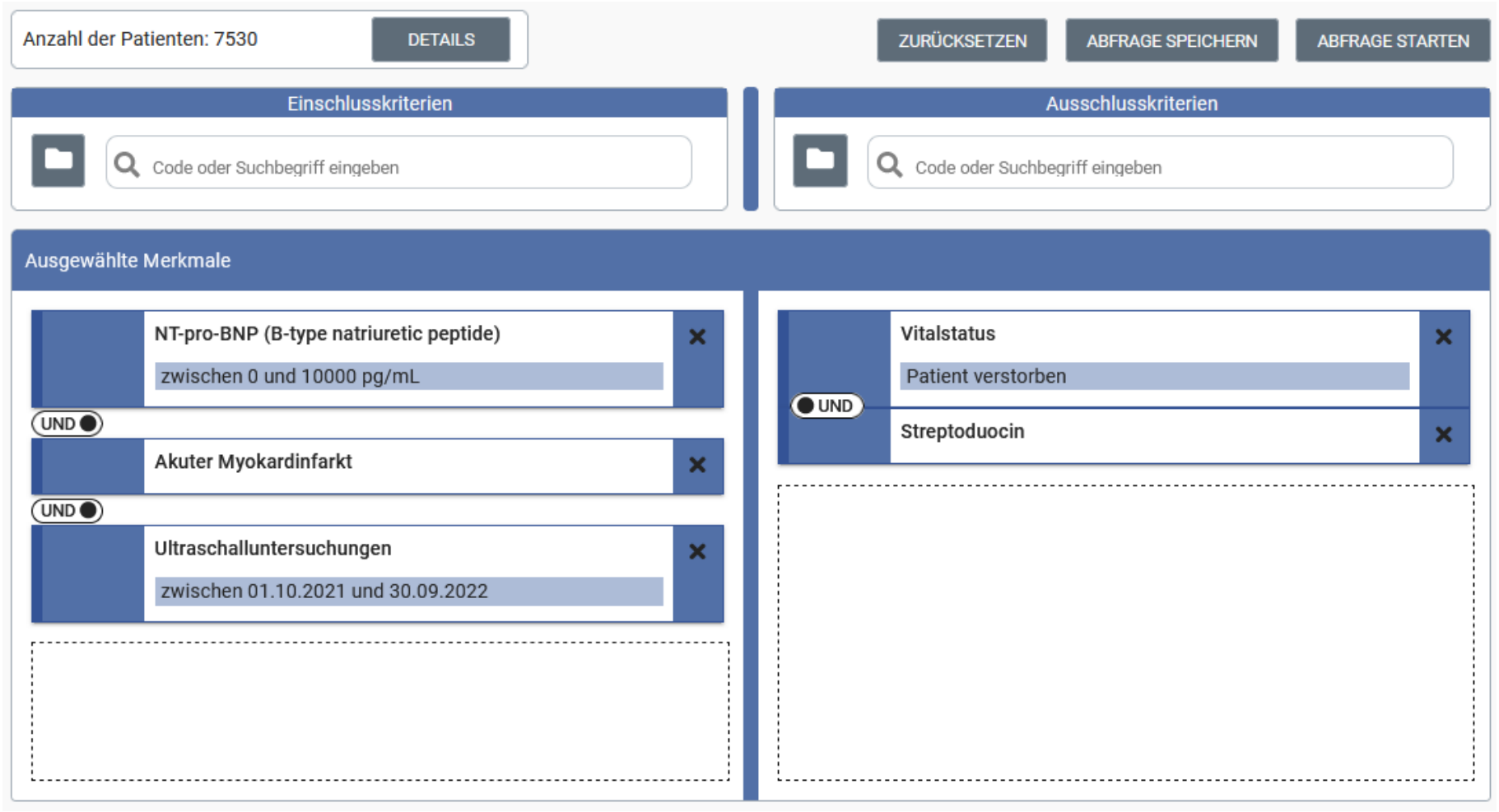


https://forschen-fuer-gesundheit.de/

Antragsverwaltung – intern



Machbarkeitsanfragen geben eine ungefähre Einschätzung der Fallzahl zurück



Datenausleitung und -bereitstellung

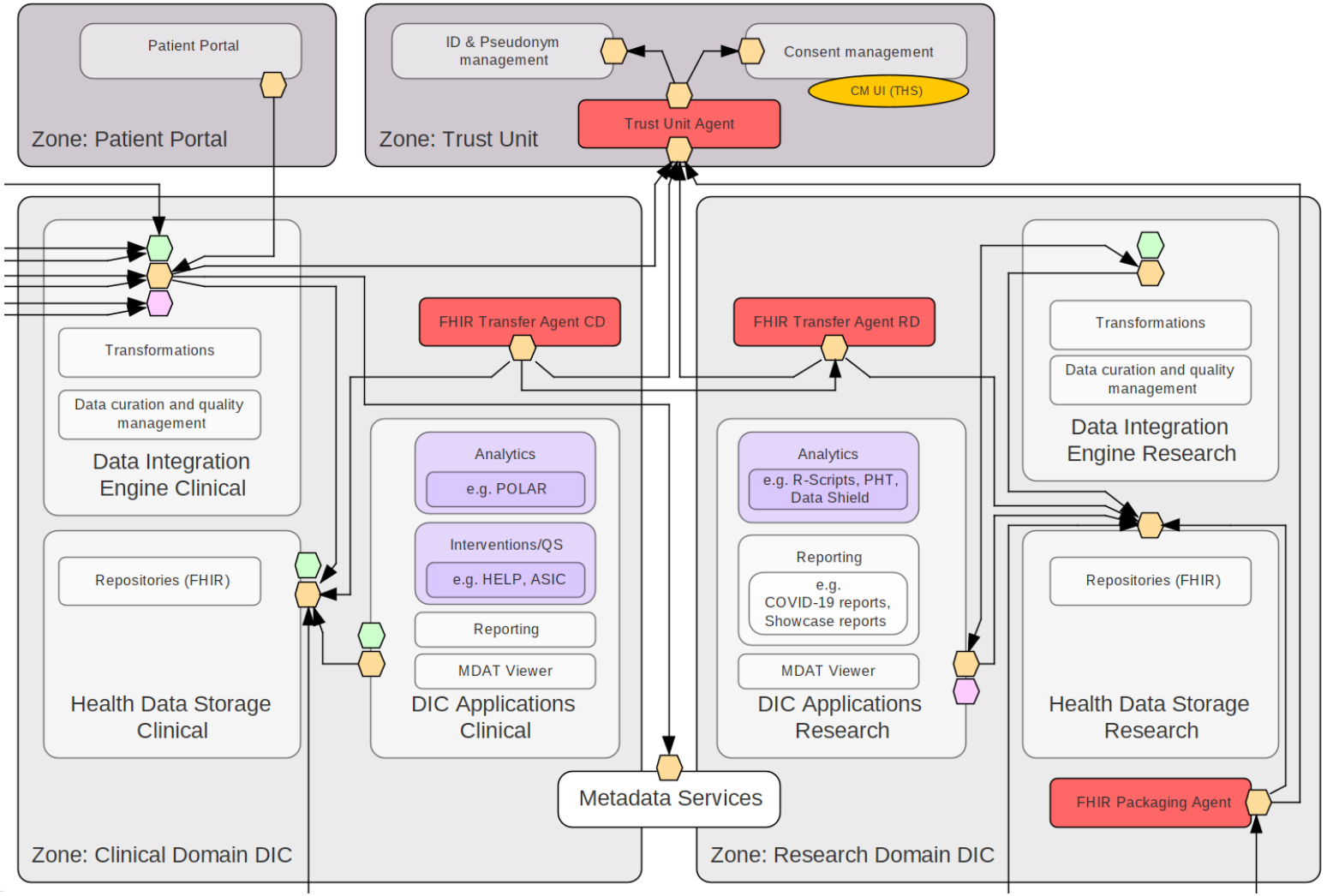
Im Falle externer Projektvorhaben ohne spezifischere Rechtsgrundlage erfolgt im DIZ:

- Prüfung der Einwilligung der Patienten (Consent)
- „Abschneiden“ von nicht-angeforderten Attributen (Datensparsamkeit)
- Ersetzen der identifizierenden Attribute (Pseudonymisierung)
- Prüfung des Re-Identifizierungsrisikos und Maßnahmen (Anonymisierung)

durch geeignete (semi-)automatische Prozesse und Werkzeuge

Trennung Clinical Domain und Research Domain in den DIZ in SMITH

- *Clinical Domain*: alle Daten aller Patienten
- *Research Domain*: Teilmenge konsentierter, pseudonymisiert Patienten und relevanter Merkmale
- Lokale unabhängige Treuhandstellen in den DIZ



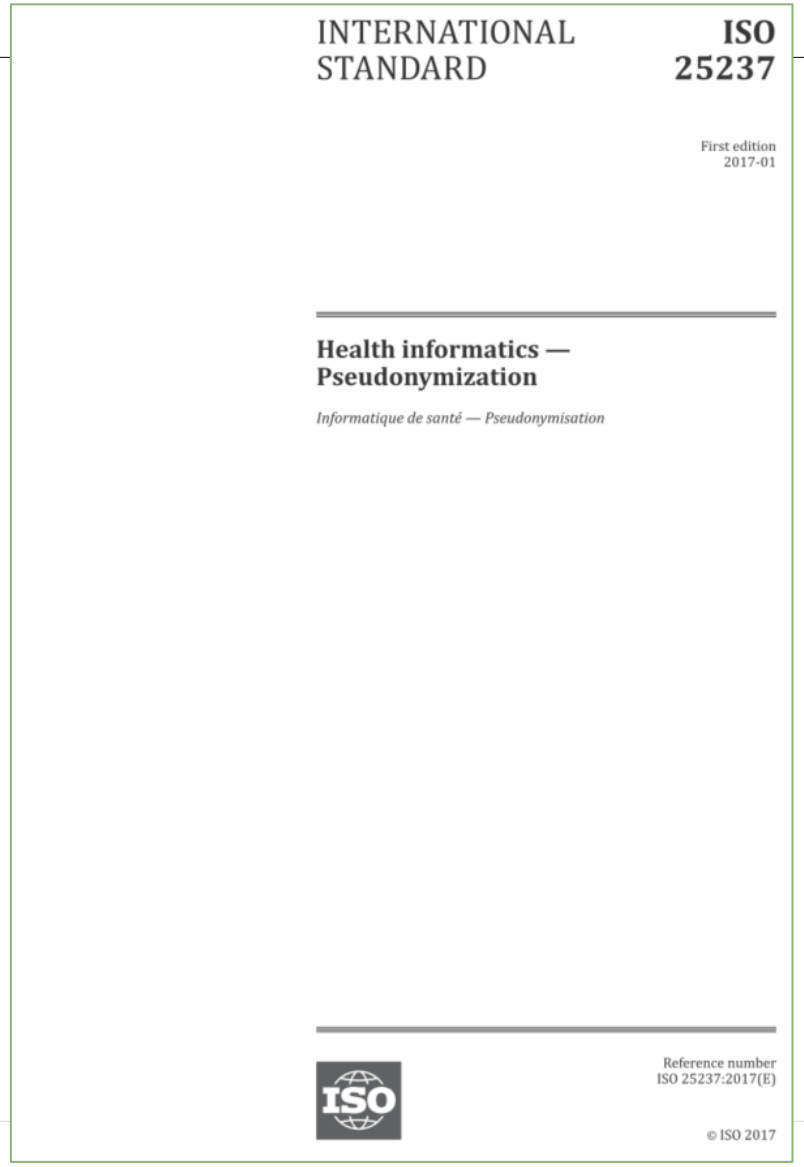
Pseudonymisierungskonzept

2.3 Pseudonymisierungsmaßnahmen (P3)

Für den Teilprozess P3 müssen die folgenden granularen Daten pseudonymisiert werden. Alle direkt identifizierenden Merkmale werden dabei entfernt bzw. ersetzt. Dies wird im FHIR Transfer Agent (CD) umgesetzt. Die Attribute und die Pflicht-Spalte beziehen sich auf das Patient-Profil des KDS-Moduls Person.

Ressource	*Attribut*	*Pflicht*	*Maßnahme*	*Kommentar / Beispiel*
Patient	identifier	Ja	Ersetzen	MPI-ID o.ä. -> TPID / SIC
	name	Ja	Ersetzen	„Max Mustermann" -> „PSEUDONYMISIERT"
	gender	Ja	Übernehmen	
	birthDate	Ja	Generalisieren	Zeit löschen und Datum auf 15. des Monats setzen (z.B. 1985-06-07 12:13:00 -> 1985-06-15)
	deceased	Nein	Übernehmen (Boolean) / Generalisieren (DateTime)	Wie birthDate
	address	Ja	Generalisieren	Straße (+ Hausnr.) entfernen, PLZ auf erste 3 Ziffern begrenzen
	maritialStatus	Nein	Übernehmen	
	multipleBirth	Nein	Übernehmen	
Alle	id / reference	Ja	Ersetzen	Hierbei geht es um die technischen Ressourcen-IDs.

ISO-Normen



INTERNATIONAL STANDARD

ISO 25237

First edition
2017-01

Health informatics — Pseudonymization

Informatique de santé — Pseudonymisation

Reference number
ISO 25237:2017(E)

© ISO 2017

INTERNATIONAL STANDARD

ISO/IEC 20889

First edition
2018-11

Privacy enhancing data de-identification terminology and classification of techniques

Terminologie et classification des techniques de dé-identification de données pour la protection de la vie privée

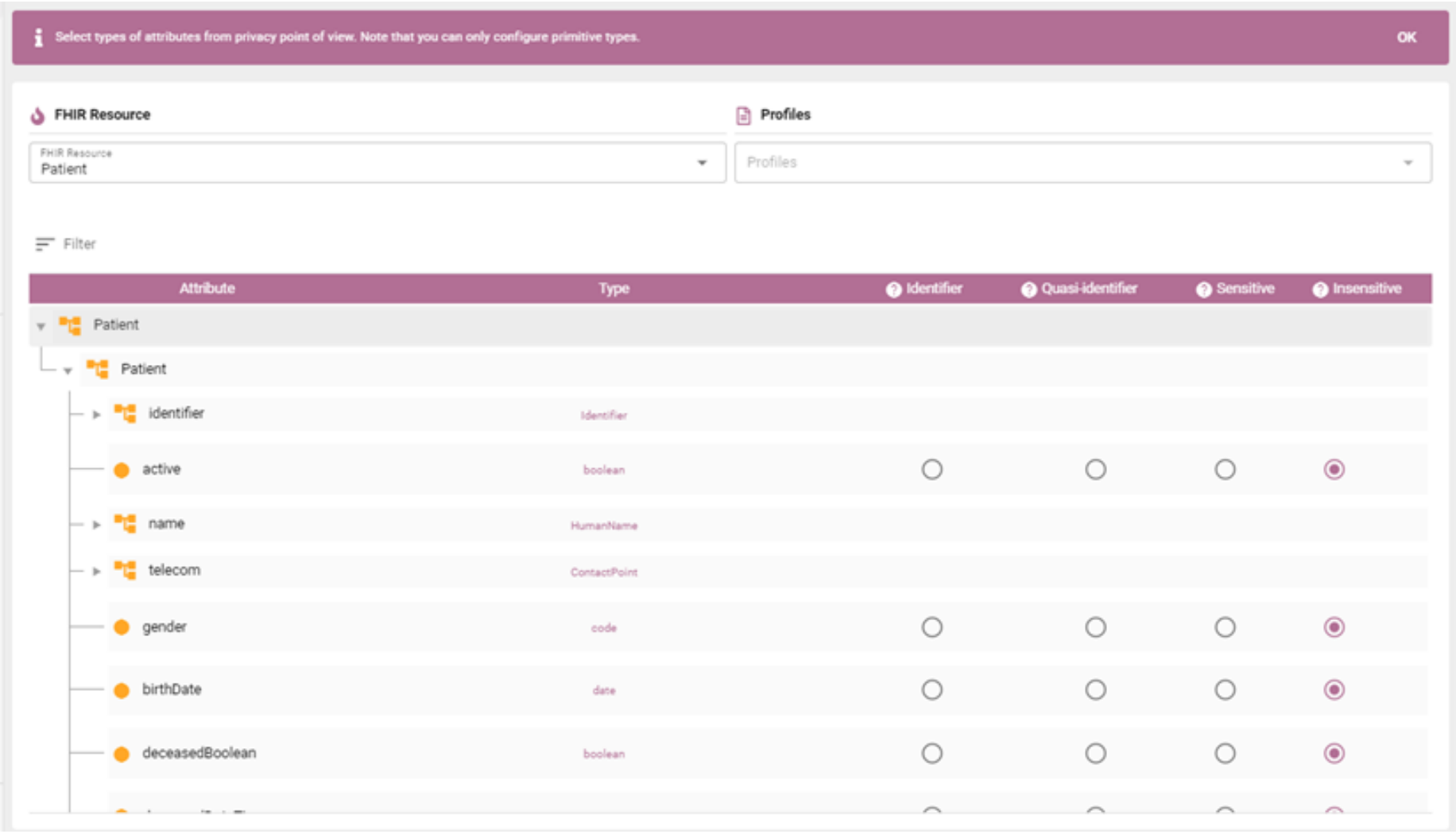
Reference number
ISO/IEC 20889:2018(E)

© ISO/IEC 2018

Beispielwerkzeug: Data Privacy Tool

- Entwickelt im Rahmen des EU-Projekts FAIR4Health
  - Arbeitet mit einer grafischen Oberfläche und auf FHIR-Datenstrukturen
  - Download: https://github.com/fair4health/data-privacy-tool/releases
  - FHIR-Server und Beispieldaten: https://github.com/fair4health/dmea-werkstatt-2022
- Auswahl von Anonymisierungsalgorithmen für einzelne Attribute einer FHIR-Ressource
  - *Pass-Through*: Attribut wird ohne Änderung gespeichert
  - *Redaction*: Attribut wird komplett entfernt
  - *Replace*: Wert wird durch Nutzereingabe ersetzt
  - *Substitution*: Ersetzung mittels regulärer Ausdrücke oder Benutzervorgaben
  - *Recoverable Subsitution*: Wert wird mittels Hash-Funktion generiert
  - *Fuzzing*: Rauschen wird mit prozentualer Vorgabe vom Benutzer hinzugefügt
  - *Date-Shifting*: Datum wird in vom Benutzer definierter Spanne zufällig verändert
  - *Generalization*: Zahlen werden nach Nutzerwusch gerundet

FHIR Ressource auswählen

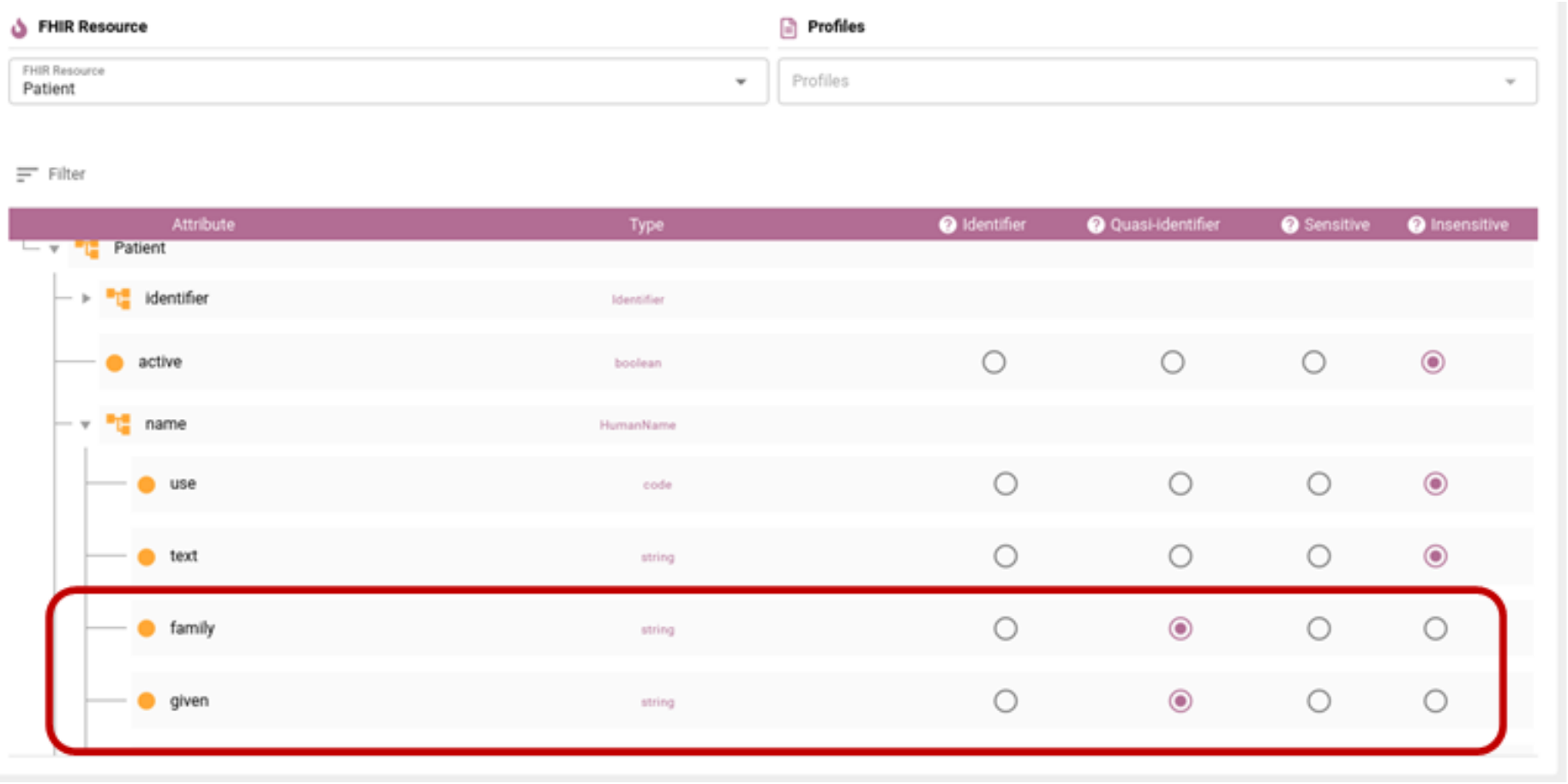


- Attribute für De-Identifikation auswählen
- Zur Auswahl stehen Identifier, Quasi-Identifier, Sensitive oder Insensitive
- Zu jeder Option erscheint eine Erklärung, wenn man mit dem Cursor über dem Fragezeichen daneben zeigt

Use Case 1: Datensparsamkeit

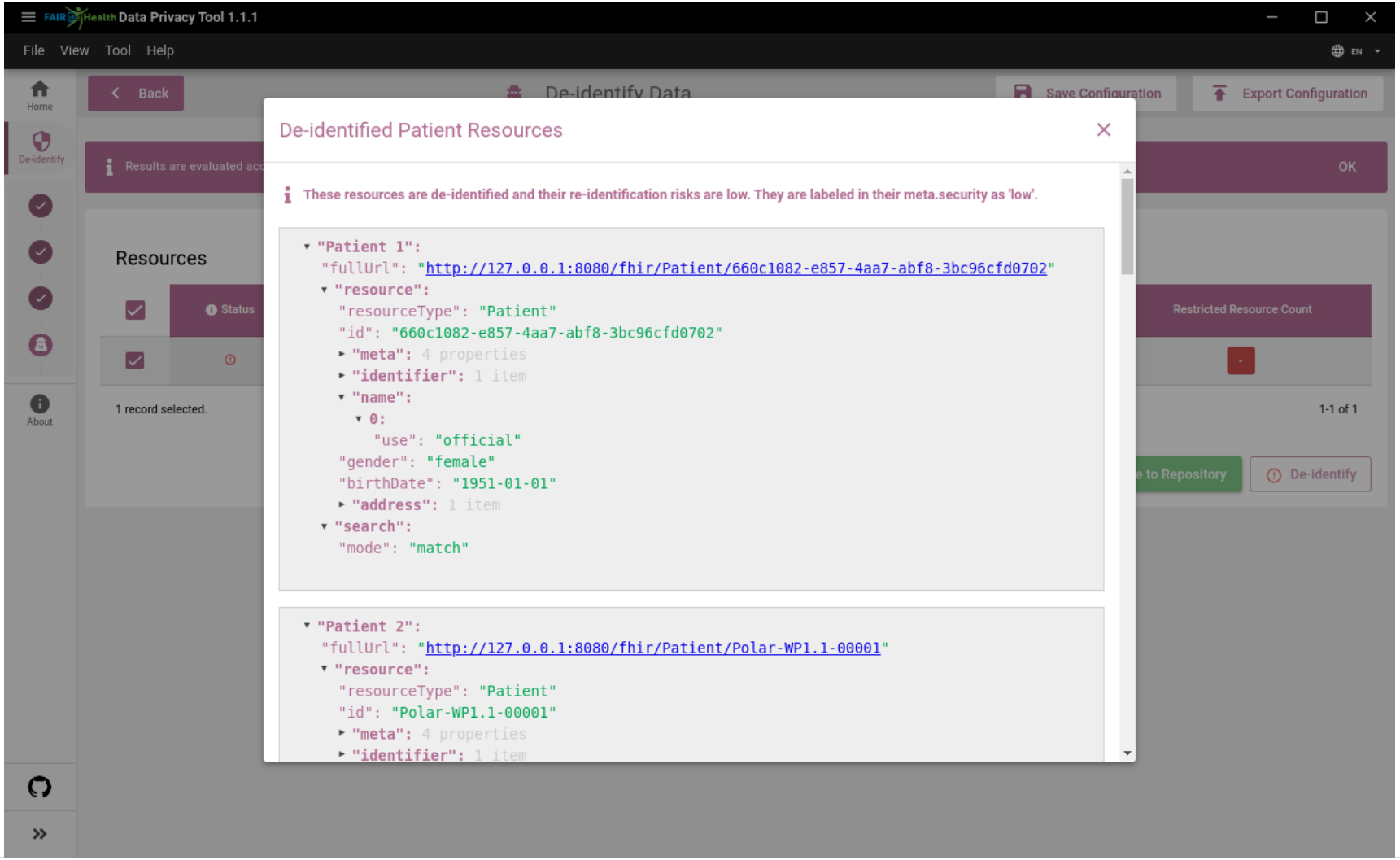
- Abschneiden von FHIR-Elementen oder -Attributen, die nicht angefordert wurden
- Anwendung bei der Weitergabe von Daten in einem DIZ an Forscher über das Forschungsdatenportal Gesundheit
- Einhalten von gesetzlichen Vorschriften zur Datensparsamkeit bei Ausgabe

Use Case 1: Attribute entfernen



- Attribute, die als Identifier markiert sind, werden automatisch entfernt
- Wenn Quasi-Identifier oder Sensitive markiert sind, muss für die Attribute Redaction ausgewählt werden

Use Case 1: Attribute entfernen

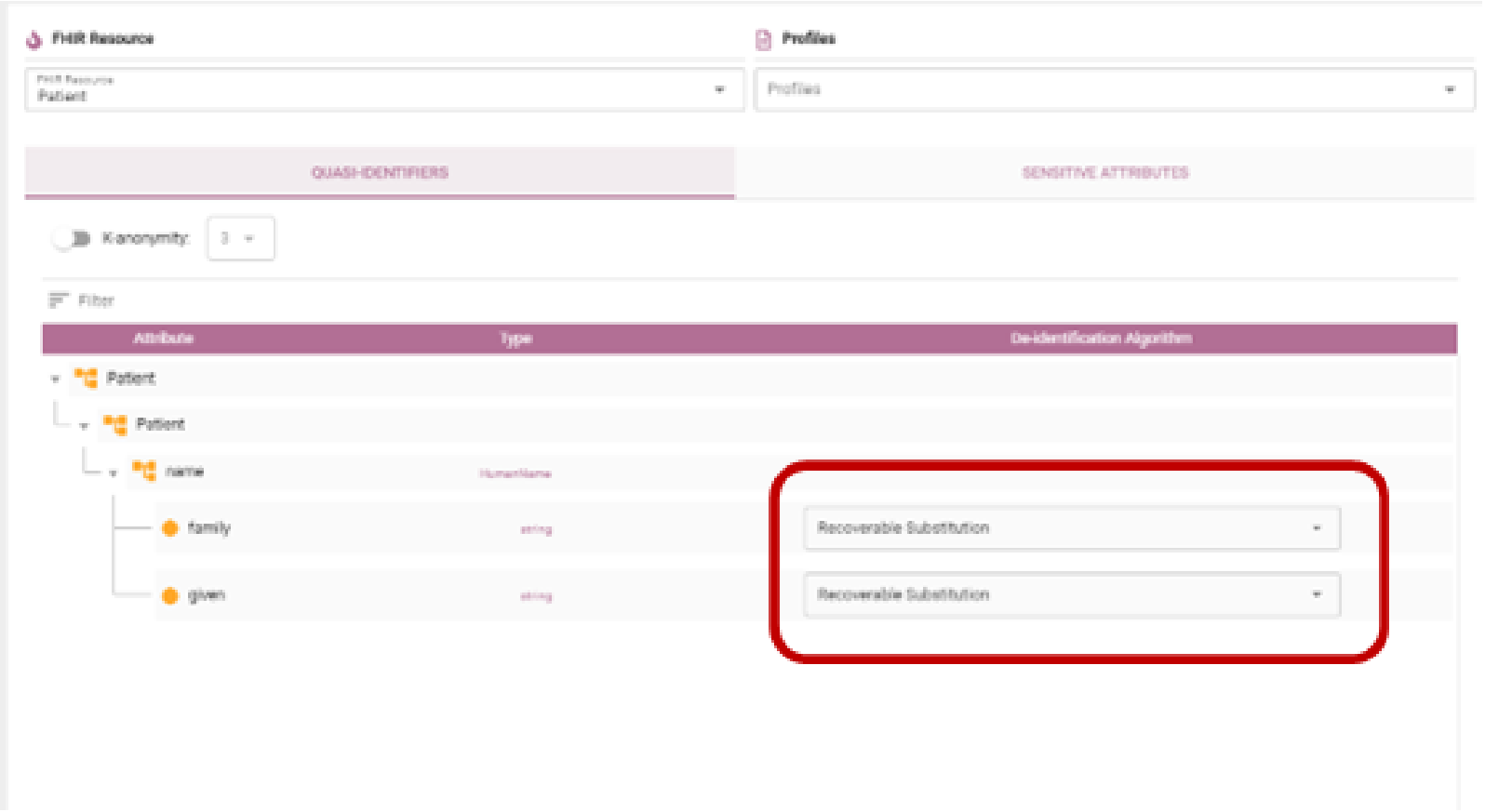


- Vor- und Nachname wurden entfernt

Use Case 2: Pseudonymisierung

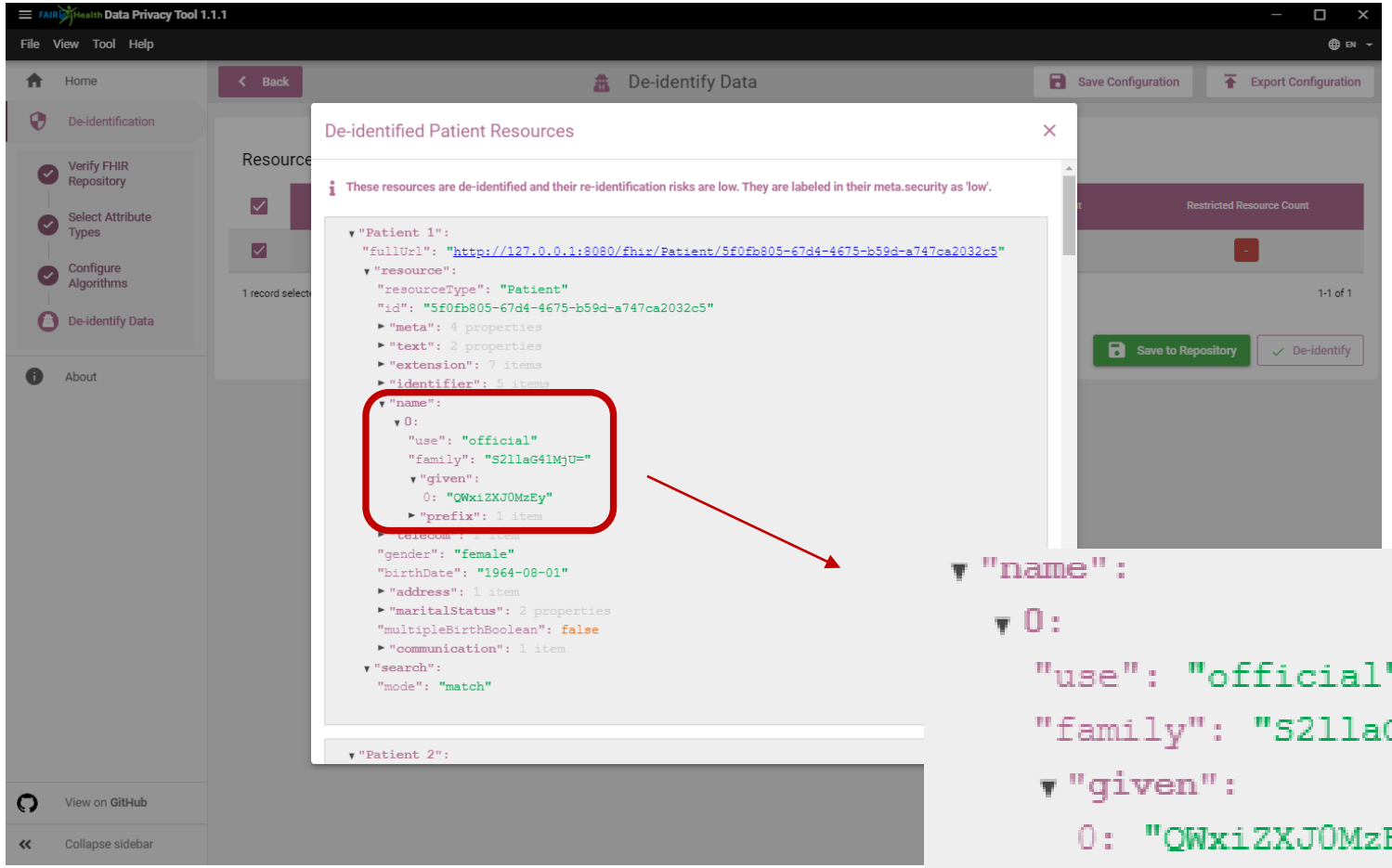
- Entfernen von identifizierenden Merkmalen wie Vor- und Nachname aus einem FHIR-Datensatz
  - Ersatz durch ein nichtsprechendes, aber eindeutiges Pseudonym
- Hintergrund: Datenausleitung in einem DIZ an lokale Forscher, die aber nicht in der Lage sein sollen, zusätzliche Daten aus dem KIS hinzuzulinken
- Zielentitäten: Primär Patientenressourcen, aber auch Fallnummern, Labor-IDs, Encounter, Observationen

Use Case 2: Pseudonymisierung von Vor- und Nachname



- *Recoverable Substitution* auswählen
- Damit können auch Identifier mittels Hashes pseudonymisiert werden

Use Case 2: Pseudonymisierung von Vor- und Nachname



- „Family name“ und „Given name“ wurden durch ein Pseudonym ersetzt

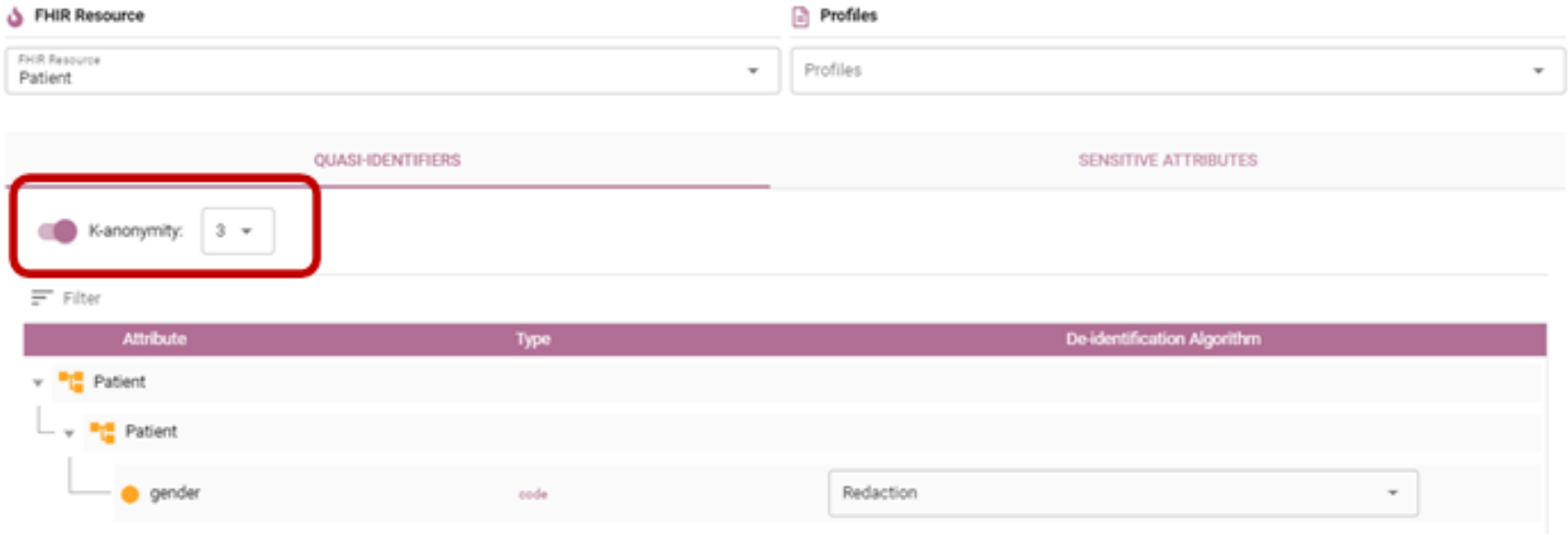
Use Case 3: Anonymisierung

- Vielzahl an wahrheitserhaltenden und nicht wahrheitserhaltenden Transformationen
- Gütemaß: k-Anonymität mit festgelegtem k

Beispiele:

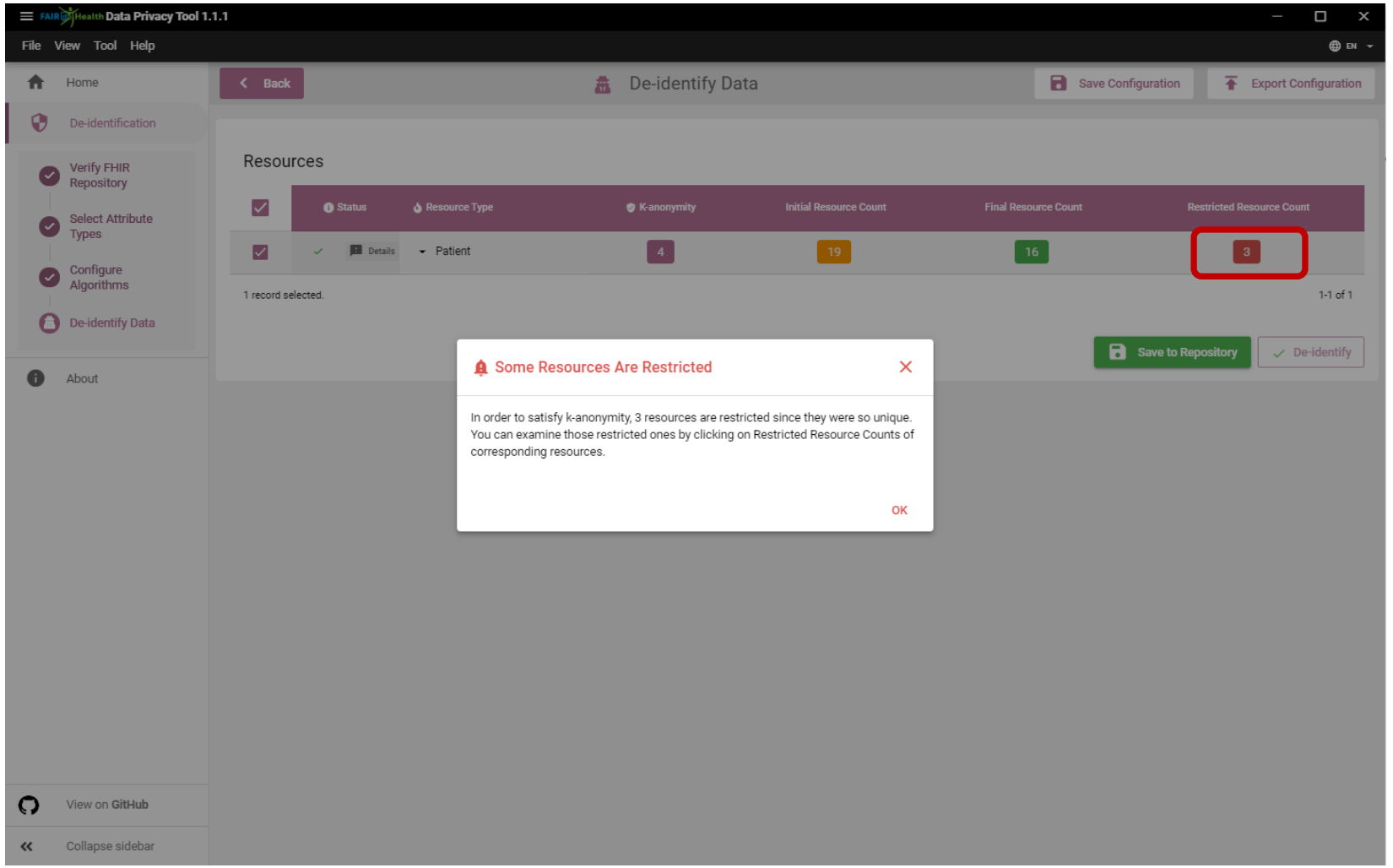
- Generalisierung
- Löschen von Zeilen (Suppression)
- Verfälschung um einen gewissen Wert
- Tauschen mit plausiblen Wert (Shift von Geburtsdatum)

Use Case 3: k-Anonymität einstellen



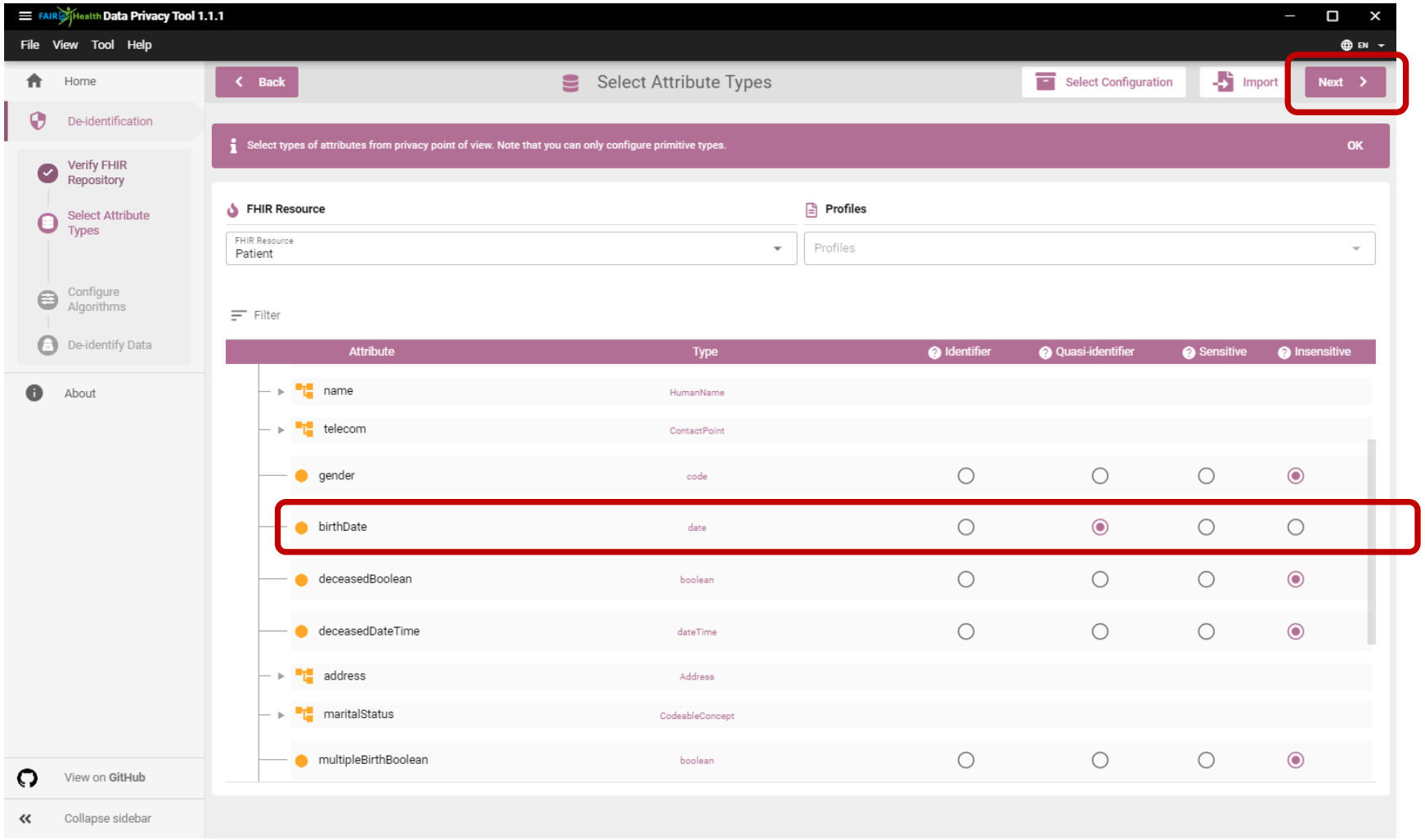
- Eine Veröffentlichung von Daten bietet K-Anonymität, falls die identifizierenden Informationen jedes einzelnen Individuums von mindestens k-1 anderen Individuen ununterscheidbar sind und somit eine korrekte Verknüpfung mit den zugehörigen sensiblen Attributen erschwert wird.
- Einstellung über Switch Button und Drop Down Menü für k-1

Use Case 3: k-Anonymität ausführen



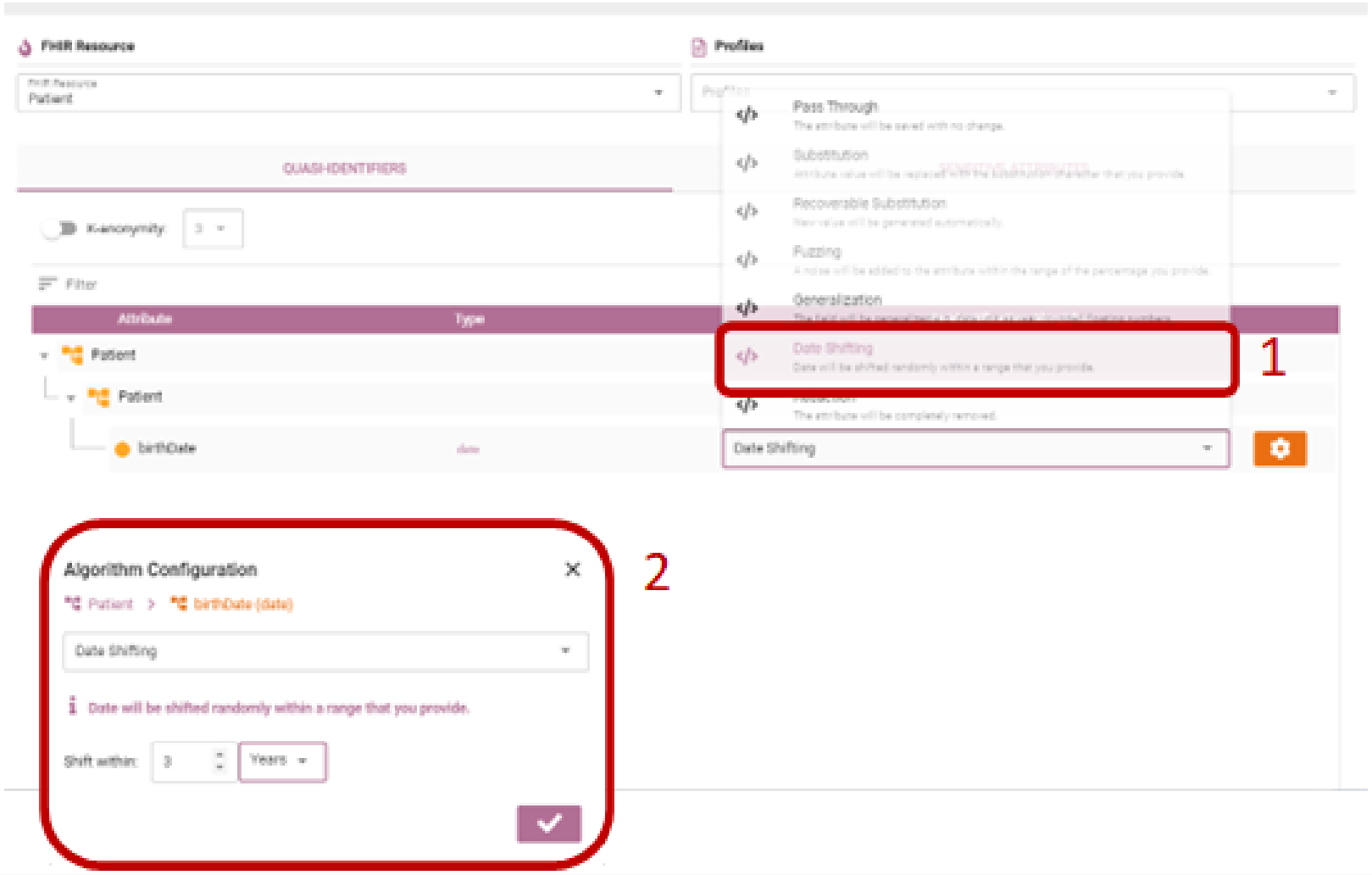
- Patienten-Ressourcen, die der geforderten k-Anonymität nicht genügen, werden unter *Restricted Resources Count* gelistet

Use Case 3: Date Shift und Generalization



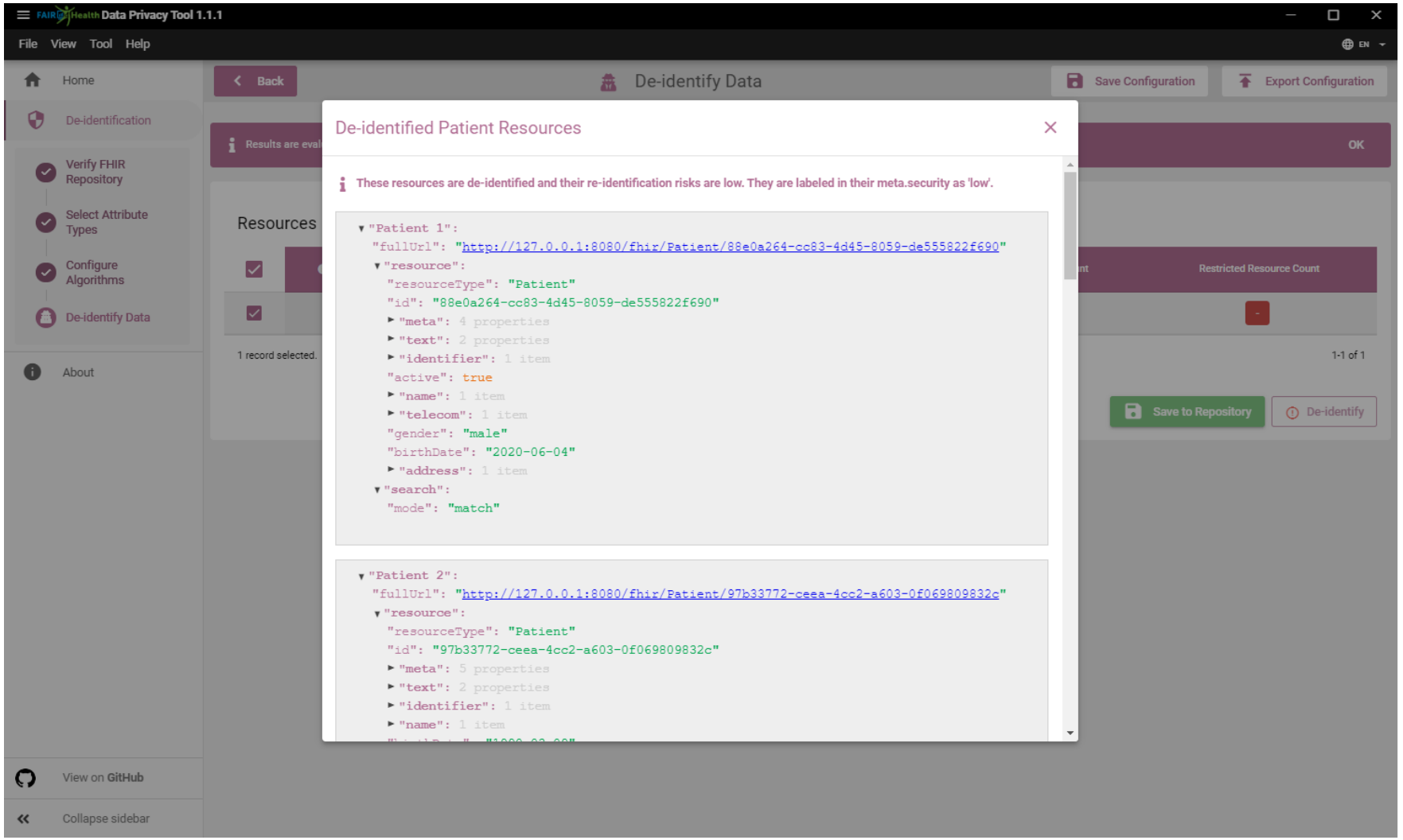
- Problem: Geburtsdatum ist in kleineren Datensätzen gern identifizierend
  - Oder bspw. Behandlungskosten in Euro
  - Man möchte aber mit den Werten rechnen
  - Kleine Abweichungen wäre verkraftbar und in der Summe unschädlich
- Geburtsdatum als Quasi-Identifier festlegen

Use Case 3: Date Shift und Generalization



- Wähle *Date Shifting* aus, um für eine festgelegte Spanne das Datum zufällig zu verändern
- Mit den orangenen Zahnrad kann man die Spanne festlegen
- Alternativ kann man mittels *Generalization birthDate* auf das Jahr reduzieren
- (nicht verfügbare Deidentifizierungsoptionen sind ausgegraut)

Use Case 3: Date Shift und Generalization

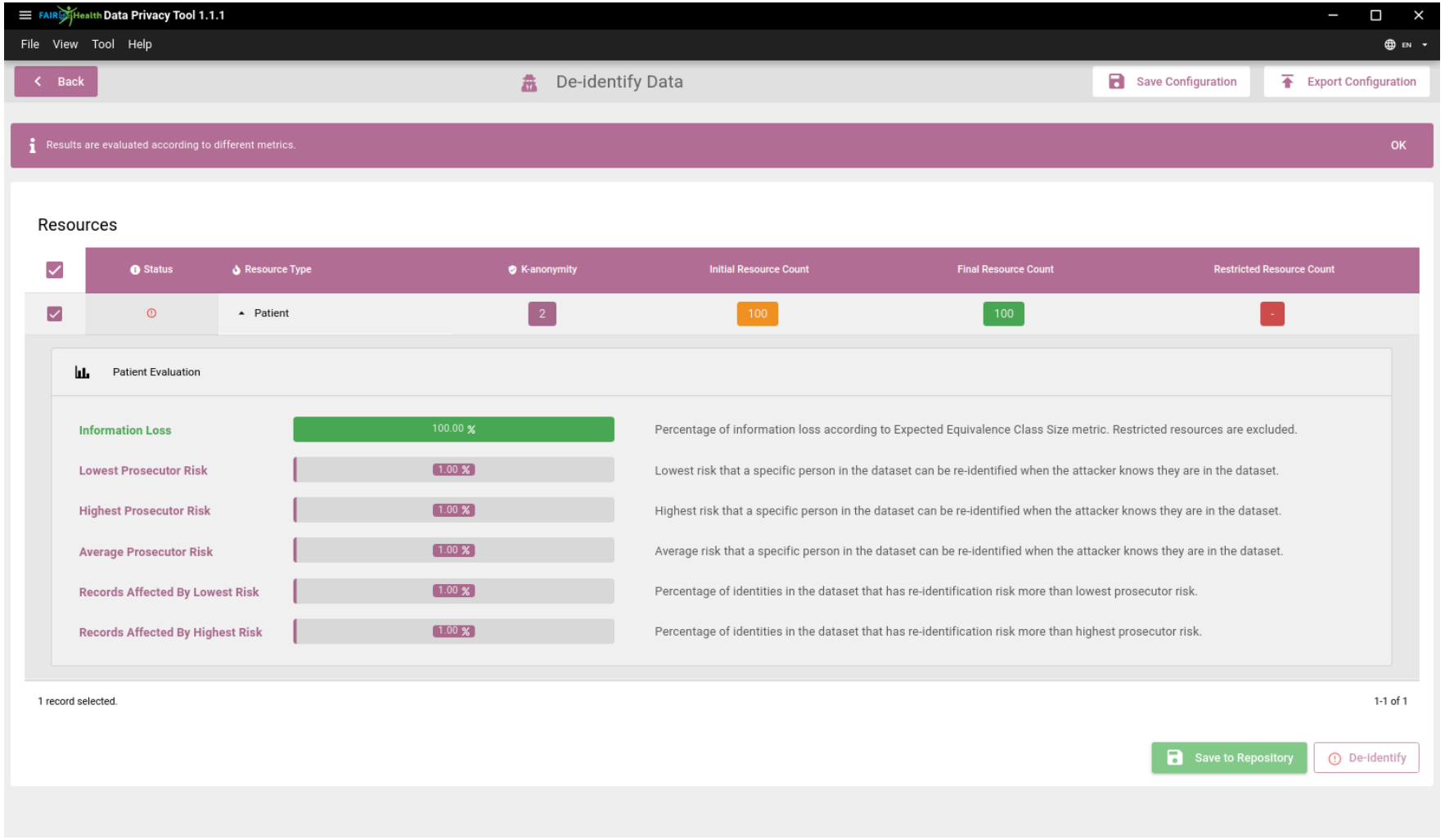


- Datumsangaben wurden durch in der Nähe liegende Datumsangaben ersetzt

Use Case 4: Risikoeinschätzung

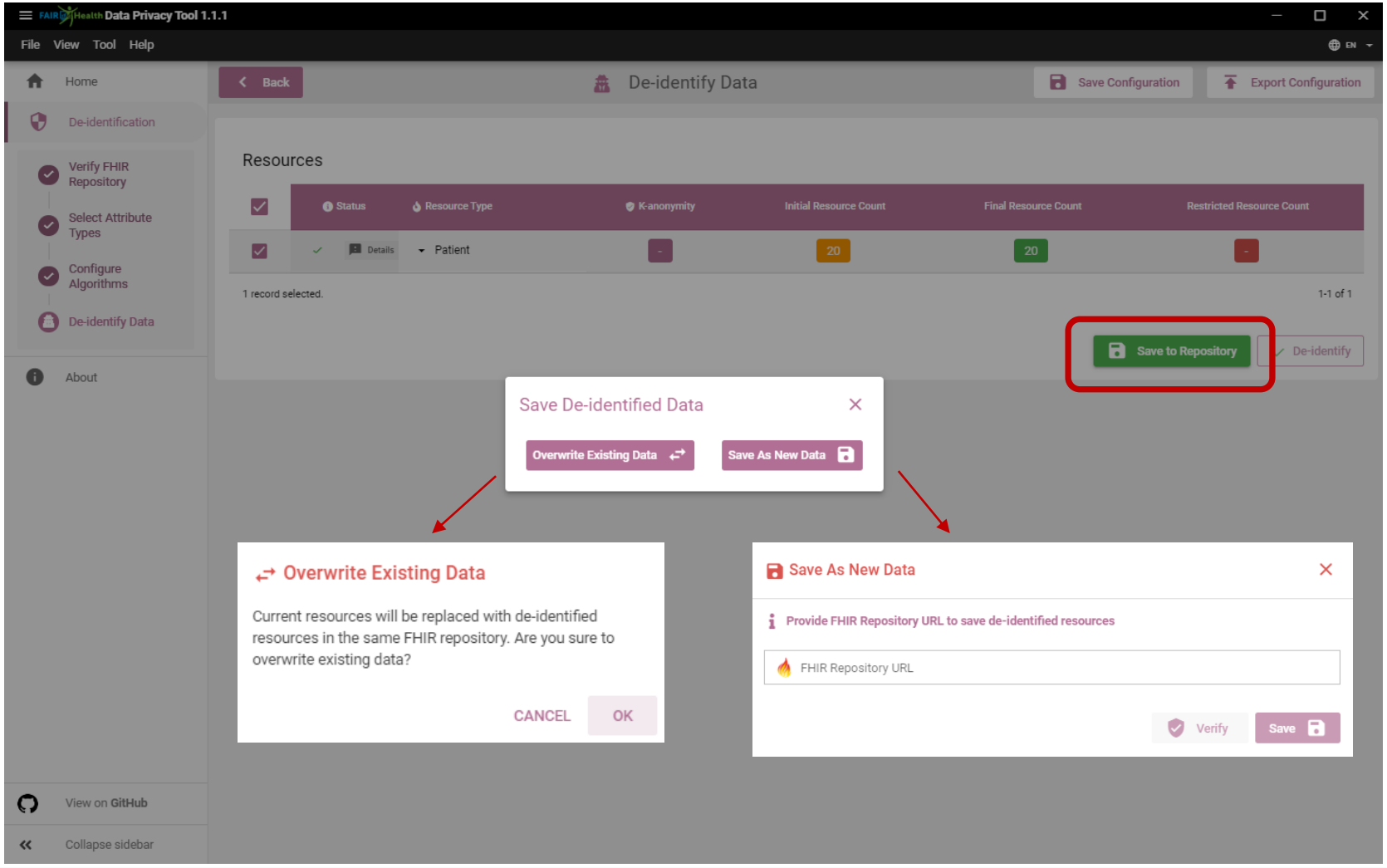
- Überblick über die Gefährdung einzelner Individuen beim Ausleiten an Dritte
- Reidentifikationsrisiko
- Hintergrund: Wenn Daten verloren gehen, ist der Zugriffsschutz nicht mehr in der eigenen Hand
- Etablierte Statistiken und Modelle für Einschätzung von typischen und extremen Fällen

Use Case 4: Risikoeinschätzung



- Information Loss
- Lowest Prosecutor Risk
- Highest Prosecutor Risk
- Average Prosecutor Risk
- Records Affected by lowest Risk
- Records Affected by highest Risk

Festschreiben der Änderungen



- *Save to Repository*, um bereits existierende Daten zu überschreiben
  - oder die anonymisierten Dateien auf einen anderen FHIR-Server zu speichern

Vielen Dank für die Aufmerksamkeit!