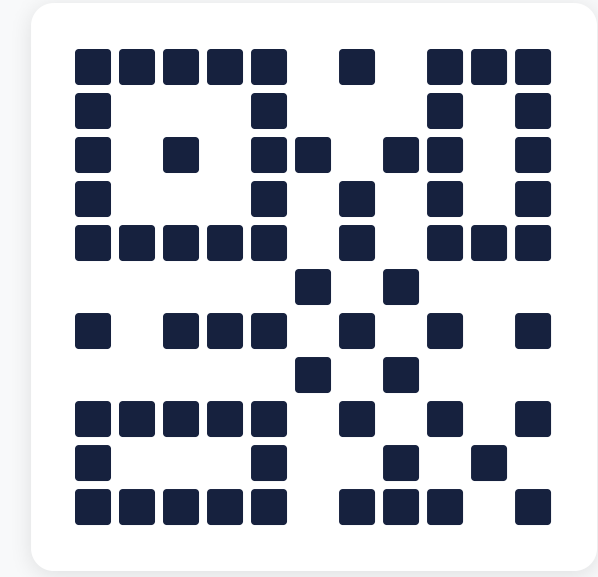
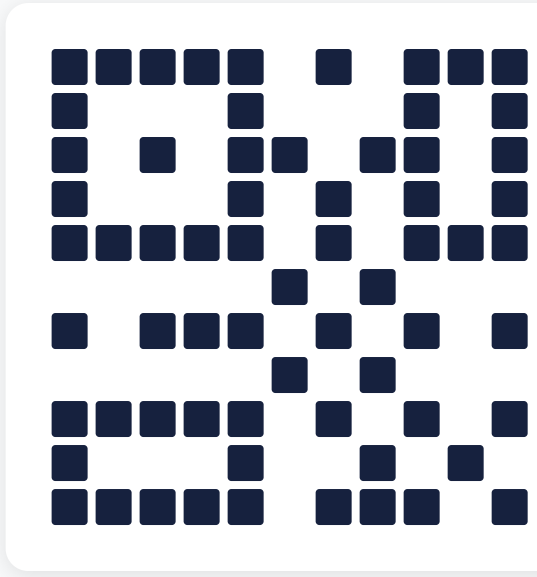




Launching OMOP Implementation on Swedish Health Data Tools and Workflows

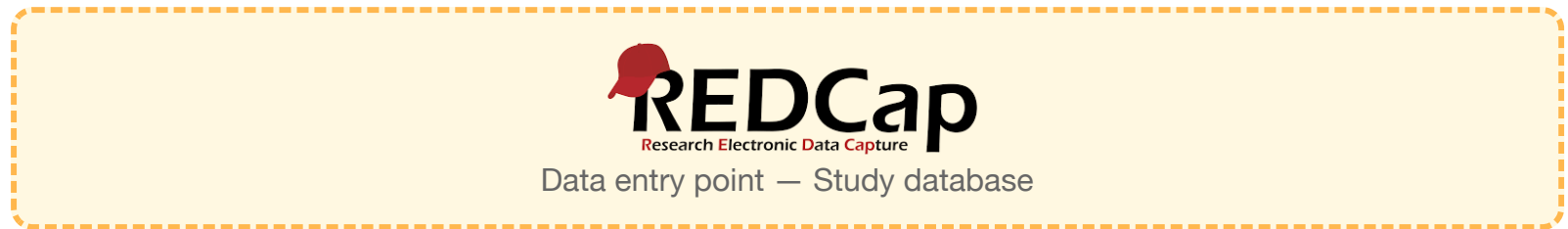
S. Rapisarda^{1,2}, J. Lorenz^{1,2}, R. Rosenberg¹, R. Fred^{1,2}, R. Caridha^{1,2}, J. Bianchi^{1,2}, I. Comarovschii¹, A. Gerhardsson^{1,2}, J. Furuhielm¹, D. Lundqvist^{1,3}, P. Östling^{1,2}

¹Karolinska Institutet · ²SciLifeLab · ³Karolinska University Hospital



Key Approach: Federated LLM-Guided Data Curation and Mapping

RESEARCHER — LOCAL ENVIRONMENT



1. Local Data Profiling

Run profiling on REDCap export: value distributions, completeness, outlier detection, database structure.

pandas-profiling / Great Expectations

CSV Profile Report Local only

OUTPUT

Data Profile & Quality Report

Distributions, completeness, anomalies

1

2. LLM-Guided Semantic Enrichment

A local LLM reads the data dictionary with domain context files and asks the researcher plain-language questions:

"The field 'hb_diag' looks like a hemoglobin blood test. What unit does your lab use — grams per deciliter or grams per liter?"

"Your form has 8 options for tumor staging. Which edition of the staging system does your study follow?"

"About 12% of patients have no performance status score. Is this only recorded for patients with advanced disease?"

Local LLM + structured markdown context

Local LLM Context files Q&A log

OUTPUT

Curated & Enriched Metadata

Semantically enriched data dictionary

2

3. Concept Matching & Review

Search OMOP vocabularies with enriched terms. Generate colour-coded review workbook for researcher approval.

OMOP Vocabulary API / Athena

Vocabulary API OMOP concepts Review workbook

OUTPUT

Review Workbook

Colour-coded candidate OMOP concept mappings

3

HANDOFF — METADATA ONLY (NO PATIENT DATA)

4

4. Submit for Expert Review

Reviewed mapping + data profile. **No patient data** — only metadata, aggregates, and proposed mappings.

Mapping Aggregates Q&A log

KI/SCILIFELAB TEAM

5. Expert Curation

Review mapping against OMOP conventions, resolve gaps. Formalize as LinkML digital object.

Mapping Platform / Athena

Curated LinkML

OUTPUT

Digital Mapping Object

LinkML schema, FAIR, versioned

5

6. ETL on Synthetic Data

Generate & test ETL pipelines on synthetic data matching researcher's distributions.

Python / SQL / Synthea

ETL SQL Synthetic

OUTPUT

Pipeline & Synthetic Data

Tested ETL code + validation report

6

DELIVER TESTED ARTEFACTS → RESEARCHER — LOCAL ENVIRONMENT

7. Run ETL on Real Data

Execute delivered ETL on patient data. Achilles/DQD quality checks. Data stays local.

Docker / Python + PostgreSQL

OMOP CDM Achilles Local only

OUTPUT

OMOP Instance

Local validated CDM database

7

PUBLISH — COMMUNITY PLATFORM

8

8. Publish Mapping to Platform

Finalized mapping digital object published on a collaborative platform, potentially maintained by the OHDSI community. Teams can browse, reuse, comment, and propose amendments.

LinkML Platform Repository

Background & Motivation

THE PROBLEM

OMOP adoption in Sweden faces barriers: restricted data access, fragmented healthcare, heterogeneous multimodal datasets, proprietary formats, and semantically poor metadata.

Current approaches require centralizing sensitive data or relying on experts with direct access to patient records.

THE GAP

The researcher understands the data best — but lacks OMOP expertise. The harmonization team has OMOP expertise — but rarely has access to data.

CORE IDEA

Data curation is the key to interoperability. Guide researchers to curate their own data at the source using LLM tools that understand the target standard.

APPROACH

Patient data never leaves the researcher's institution. Only **metadata, data profiles, mappings, and pipelines** are exchanged. The researcher runs all data-touching tools locally; the central harmonization team reviews mappings and provides tested ETL pipelines built against synthetic data.

A local LLM receives domain knowledge through versioned context files — not fine-tuning. For example, oncology: OMOP Oncology Extension, staging conventions, cancer vocabularies (NAACCR, ICD-O).

Different domain = different context pack. Same tool, same workflow.

SCOPE

TEF-Health demonstrator projects, five oncology studies at Karolinska Institutet: BioBladder, FOCUSE, PreDDLung, SprintR, PREDIX HER2. Potential addition of EHR data from region Stockholm in later phases.

ASK ME!

Tools & Platform

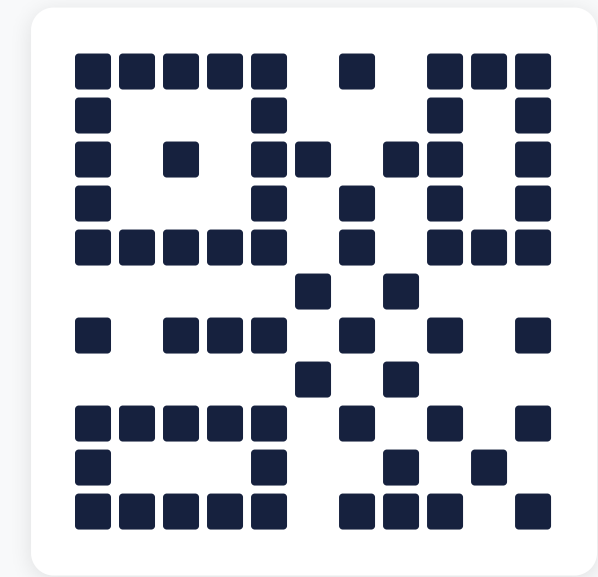
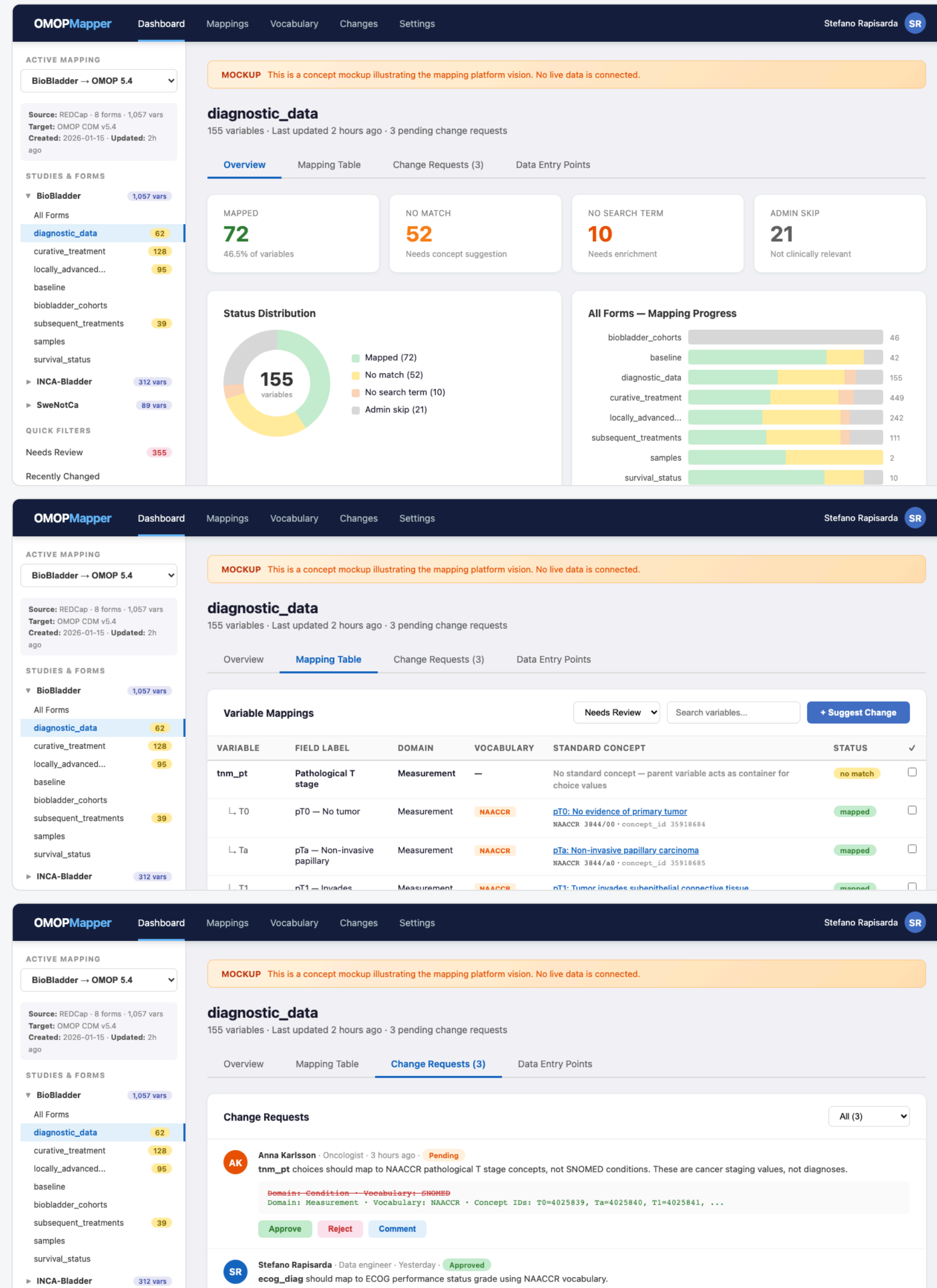
AI Curation Tool

An LLM-powered tool that guides researchers through data curation without requiring OMOP expertise. The tool profiles the source data, identifies gaps and ambiguities, asks targeted questions in plain language, and produces standardized mappings ready for expert review.

- 1 Data Profiling**
Distributions, completeness, outliers
- 2 Quality Assessment**
Value ranges, integrity, anomalies
- 3 Gap Analysis**
Missing metadata, unmapped vocabularies
- 4 Researcher Q&A**
Plain-language questions on data
- 5 Semantic Enrichment**
OMOP domains, search terms, units, data entry points, data collection workflows
- 6 Automated Mapping**
Concept matching, propose IDs

Mapping Platform (concept)

A collaborative repository for OMOP mapping digital objects — inspired by code-hosting platforms. Teams browse existing mappings by study and domain, review variable-level details, and propose changes through a structured review process. New oncology studies can start from an existing mapping rather than from scratch.



SciLifeLab

TEF-Health Sweden

