ETL FOR OMOP CDM FROM SYNTHEA

NABC ACTION PLAN

TEAM OMOPSYNC

LOF TA - MERLIN SIMOES
AUM SATHWARA (A20543213)
BJOERN SAGSTAD (A20557181)
HARNEET KAUR (A20548613)
VISHNU SHANMUGAVEL (A20561323)

We have all used ChatGPT...

What if structuring Healthcare Data was as simple as prompting a LLM?

"Hey, here's my CSV file — can you convert it to OMOP CDM."

And just like that, the **“AI”** triggers the backend pipeline that handles everything — no manual scripting.

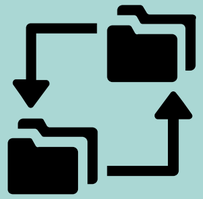
Hook

Healthcare already produces ≈ 30 % of the planet’s data, and volume is set to leap from 2.3 ZB in 2020 to 10.8 ZB by the end of 2025 (36 % CAGR); **traditional, hand-built ETL pipelines simply cannot keep up with that growth.**

Source

2024 HIMSS survey found that 47 % of healthcare leaders are **dissatisfied with their organizations’ data quality**, citing budget limits and **manual workflows** as the top barriers.

Source



Fewer than half of U.S. hospitals (43 %) routinely exchange and integrate external patient information, and roughly 30 % remain not fully interoperable, **hampering cross-institutional studies**

Source

Need

Researchers need a simple tool to convert diverse healthcare datasets into the OMOP CDM format.

AI-powered, scalable ETL pipelines that standardize, clean, validate, and index diverse healthcare data into the OMOP CDM.

What is OMOP CDM?

Standardized format for healthcare data, making it ready for analysis, sharing, and AI.

Ease of Coversion

A simple prompt to convert to OMOP CDM

Data Quality Verification

DQD shows data passes 98% checks

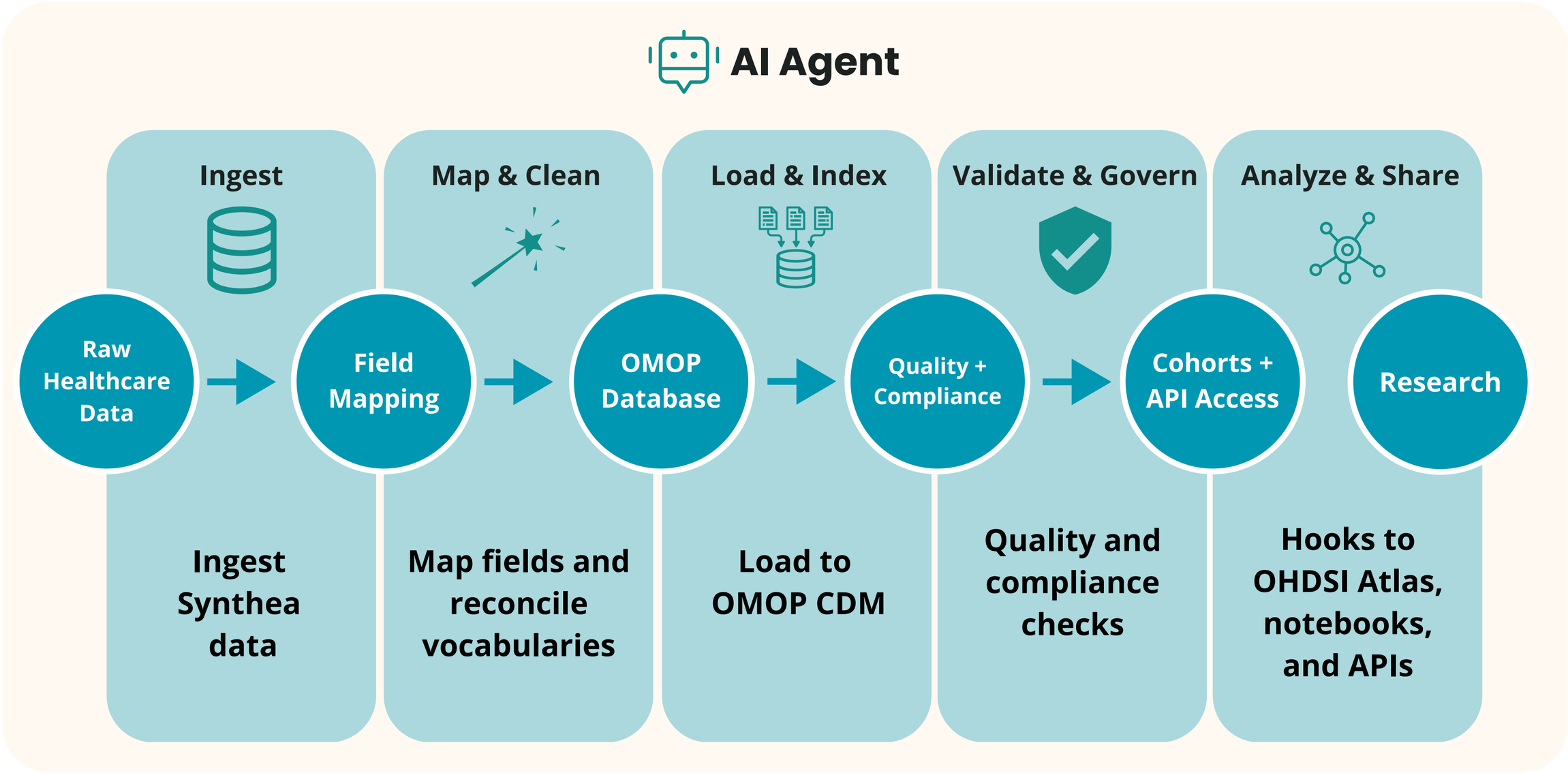
Support for Analysis

Run queries on data with prompts

Cross-Institutional Interoperability

Make data shareable across organizations.

Approach


AI Agent
Ingest
Map & Clean
Load & Index
Validate & Govern
Analyze & Share
Raw Healthcare Data
Field Mapping
OMOP Database
Quality + Compliance
Cohorts + API Access
Research
Ingest Synthea data
Map fields and reconcile vocabularies
Load to OMOP CDM
Quality and compliance checks
Hooks to OHDSI Atlas, notebooks, and APIs

BENEFITS/COSTS

S

STRENGTHS

- AI-Triggered ETL Automation
- OMOP + OHDSI Standardization
- AI Assisted queries, analysis
- Understands natural language queries

W

WEAKNESSES

- Need for Regular Vocabulary Updates
- Create custom ETL pipelines for different datasets

O

OPPORTUNITIES

- Rising Demand for Interoperable Research
- Expansion into Real-World Data and FHIR
- Academic, Clinical, Commercial Integration
- Integration of AI analysis

T

THREATS

- Competition from Established ETL Vendors
- Changing Compliance Regulations (HIPAA, GDPR)
- Institutional Resistance to New Pipelines

Market Competition

IQVIA™

IQVIA™ OMOP Converter

Customized SQL-based solutions to convert your data to OMOP

The Hyve

The Hyve's Delphyne

Open-source OMOP ETL. Efficient, yet lacks adaptability for custom data needs.

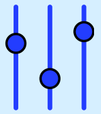
In House

Inhouse ETL Pipelines

Custom academic tools built in-house at universities / hospitals / research institutions

Our Edge

Minimal Setup



Fully Customizable

Transparency

Shallow Learning Curve

Data Privacy

Risk Mitigation

Data Integration

Low **High**

Mapping diverse healthcare datasets to the OMOP CDM format.

- Build detailed integration strategies and test data quality.

Regulatory Compliance

Low **High**

HIPAA non-compliance risks.

- Engage legal teams early, anonymize data, ensure HIPAA adherence.

Vocabulary Management

Low **High**

OMOP vocabulary inconsistencies.

- Automate updates using OHDSI tools and monitor regularly.

Technology Scaling

Low **High**

Challenges scaling ETL for large datasets.

- Optimize ETL scripts to reduce the risk of bottlenecks

Action Plan

Understand the Synthea data and how it can be mapped to OMOP CDM

Working ETL pipeline converting Synthea to OMOP, with A.I. agent integration.

Run scripts and visualize results of queries on the population health.

Thank You!

TEAM OMOPSYNC