

What does it take to deliver a production-ready OMOP ETL? Lessons from a hands-on workshop on the Swedish register data.

Authors: Valentina Giunchi¹, Michele Fusaroli¹, Pablo Maldonado¹, Anders Sundström¹, Huiqi Li², G. Niklas Norén¹, Judith Brand¹

¹ Uppsala Monitoring Centre, Uppsala, Sweden

² University of Gothenburg, Gothenburg, Sweden

Background: The OMOP Common Data Model (CDM), developed by the OHDSI community, offers a standardized way of representing diverse observational health data, enabling scalable, reproducible research across institutions¹. However, many organizations remain hesitant to adopt the OMOP CDM because the Extract-Transform-Load (ETL) process is perceived as highly resource intense, requiring substantial time, expertise and sustained effort²

We have recently initiated the mapping of Swedish health register data to the OMOP CDM for drug safety research. Here we report our experience, identifying key competencies, preparation steps, and collaborative practices that enable efficient ETL design and implementation, with a focus on capacity building within the OHDSI community.

Methods: We organized a one-week, hands-on workshop focused on advancing the mapping of Swedish register data to the OMOP CDM, preceded by substantial preparatory work. Planned activities included refining the ETL design using Rabbit-in-a-Hat³, performing vocabulary mappings, improving the SQL-based implementation, and conducting iterative data quality assessment. Observations were recorded throughout the process, focusing on preparation requirements, key competencies needed, collaboration dynamics, and barriers encountered during ETL design and implementation.

Results:

We identified key dimensions critical to successful ETL development for new adopters:

1. Preparation and data readiness:

Pre-workshop preparations included review of OMOP CDM specifications, mapping conventions, and OHDSI training materials, as well as downloading relevant vocabularies and profiling the source data using WhiteRabbit⁴. A functional and secure technical environment (PostgreSQL database, R orchestrator calling SQL scripts, OHDSI HADES ecosystem⁵) was established, and ARES⁶ was deployed to support data characterizations and quality review. Initial mapping to core CDM tables, cohort sampling to reduce computation time, and early ETL script drafting and testing enabled efficient use of the

workshop time. In our implementation, data curation and cleaning were performed within the ETL process to enhance transparency and traceability (**Figure 1**).

2. Key competencies:

The workshop involved seven participants with complementary expertise, including data and domain experts, epidemiologists/biostatisticians, and data engineers, supported by two experienced OHDSI contributors: one OHDSI leader (Patrick Ryan) and one experienced OMOP user specializing in Swedish register data.

This composition enabled coverage of three key competency areas required for successful ETL development: (i) data/domain expertise to correctly interpret the source data structures and semantics; (ii) CDM and vocabulary expertise to ensure appropriate mapping to standardized concepts; and (iii) technical expertise for efficient implementation and infrastructure setup.

3. Facilitators:

Enabling factors included version control systems, shared workflows, and the presence of experienced OHDSI contributors, which facilitated knowledge transfer and alignment with community standards.

Although team members had distinct areas of expertise, progress relied on continuous interaction and shared problem-solving rather than strictly sequential or isolated tasks.

4. Barriers:

Key challenges included ambiguities in source data interpretation, mapping source data vocabularies to standardized vocabularies, and the need for iterative refinement of the ETL logic. Documentation and ensuring portability of the ETL across different data holders and technical environments were also identified as time-consuming and non-trivial tasks.

5. Reproducibility and sharing:

There were few publicly available ETLs for our source data. This highlights the importance of sharing ETL implementations with clear documentation of mapping decisions and vocabulary usage. A key outcome of this work is the commitment to openly share the developed ETL and to foster more coordinated open-source efforts within the OHDSI Sweden community.

6. End-to-end implementation:

The workshop, together with the preparatory work, resulted in substantial progression toward a complete end-to-end pipeline, including ETL design, implementation, and initial data quality assessment (**Figure 1**).

Conclusion: This work suggests that OMOP ETL development can be effectively initiated through a structured, workshop-based approach, provided that key competencies, preparatory steps, and collaborative practices are in place. Our experience highlights that ETL development is not solely a technical task, but a socio-technical process requiring coordinated expertise and iterative refinement.

Importantly, while the core activities were concentrated in a one-week workshop, they built upon substantial preparatory work conducted beforehand. The involvement of experienced OHDSI contributors played a critical role in aligning practices with community standards.

This work suggests that lowering barriers to OMOP adoption requires not only sharing ETL code, but also providing supporting materials, including documentation, training resources, and guidance on infrastructure setup. By sharing both our experience and resulting ETL resources, we aim to show that the perceived complexity of ETL should not deter institutions from adopting the OMOP CDM, and that structured, collaborative approaches can make this process feasible.

References:

1. OMOP Common Data Model. Accessed May 29, 2026. <https://ohdsi.github.io/CommonDataModel/>
2. Informatics OHDS and. *Chapter 6 Extract Transform Load | The Book of OHDSI*. Accessed May 29, 2026. <https://ohdsi.github.io/TheBookOfOhdsi/>
3. Rabbit in a Hat. Accessed May 29, 2026. <https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html>
4. WhiteRabbit for ETL design – OHDSI. Accessed May 29, 2026. <https://www.ohdsi.org/analytic-tools/whiterabbit-for-etl-design/>
5. Schuemie M, Reps J, Black A, et al. Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. In: Bichel-Findlay J, Otero P, Scott P, Huesing E, eds. *Studies in Health Technology and Informatics*. IOS Press; 2024. doi:10.3233/SHTI231108
6. Home | Ares. Accessed May 29, 2026. <https://ohdsi.github.io/Ares/>

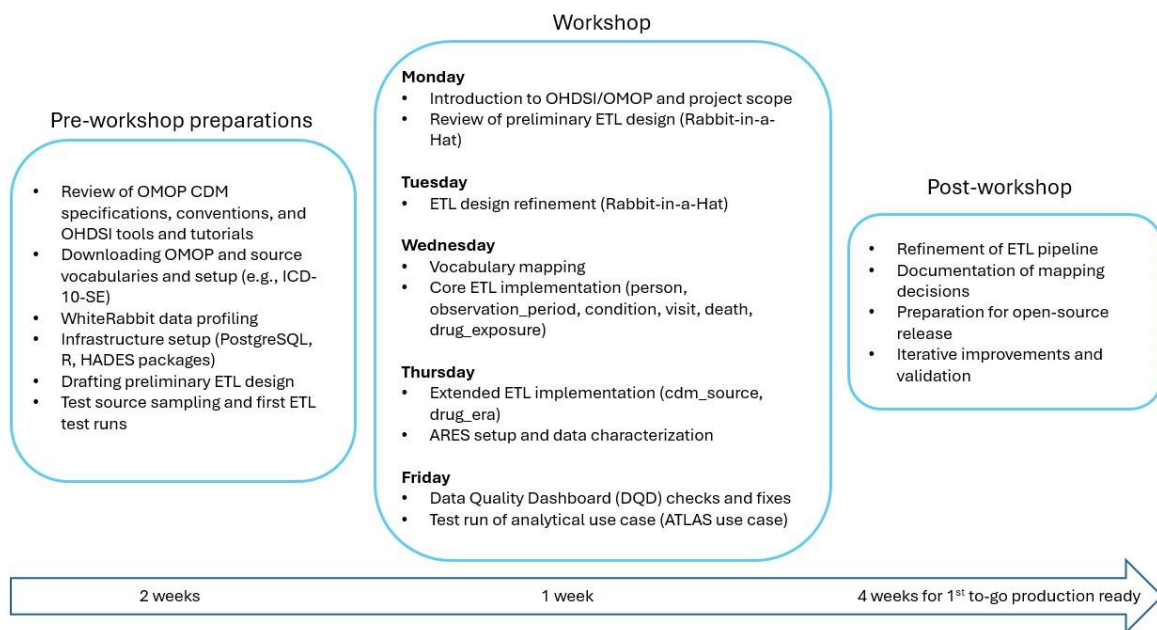


Figure 1 – Timeline of pre-workshop preparation, hands-on workshop activities, and post-workshop follow-up