Large-Scale Extraction and Standardization of Primary Care Vaccination Data to OMOP CDM in Wallonia

Building a population-scale, interoperable vaccination dataset from general practice records

Background: The secondary use of primary care data is essential for observational research and public health surveillance. In Belgium, general practitioner electronic health records (sumEHR) provide broad population coverage but remain heterogeneous across systems and coding practices. This study describes the large-scale extraction, standardization, and validation of vaccination data from primary care in Wallonia, mapped to the OMOP Common Data Model (CDM).

Method

Vaccination data were extracted from sumEHR collected via the Health Network Réseau Santé Wallon (RSW), covering approximately 1.5 million patients between 2012 and 2025, for a total of nearly 7 million records.

A dedicated ETL pipeline was developed to map the data to the OMOP CDM, including filtering, validation, and terminology standardization. Vaccination identification followed a hierarchical strategy using Belgian drug packaging codes (CNK), ATC classification, and KMEHR (Kind Messages for Electronic Healthcare Record) terminology which is a Belgian standard for medical data.

Approximately 96.8% of records were successfully mapped through this process.

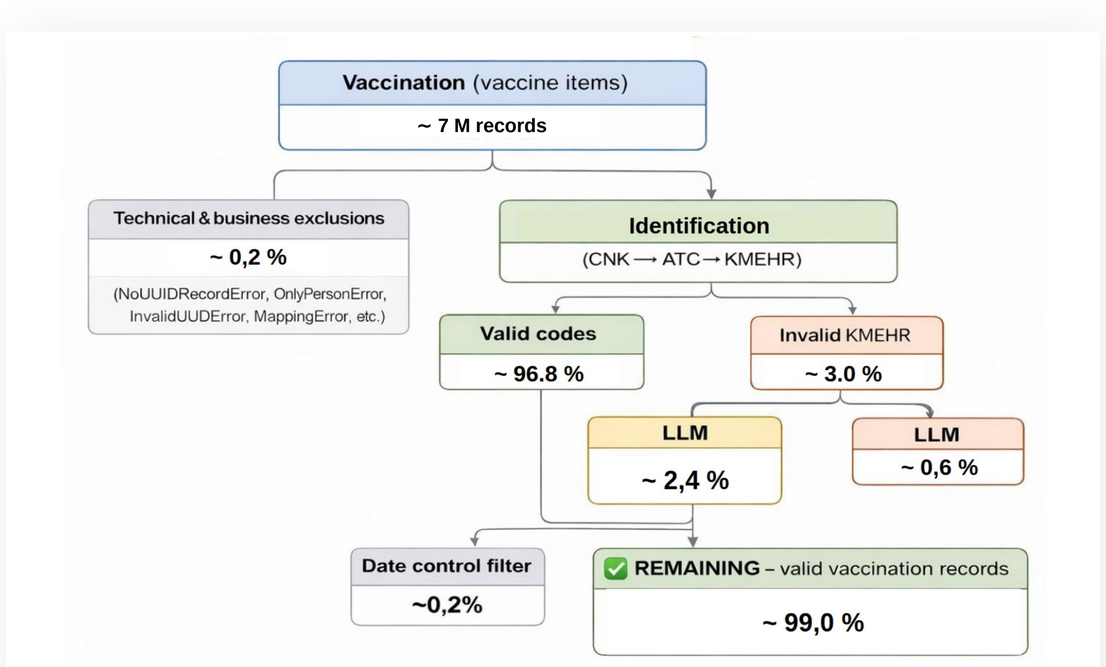
For the remaining unresolved records (~3%), LLM-assisted methods were applied as a complementary step, allowing recovery of most entries while maintaining quality standards. A final temporal validation step ensured consistency of vaccination dates.

The pipeline combines deterministic mapping strategies with probabilistic recovery methods to maximize completeness while preserving data reliability.

It is designed to be scalable and reproducible across other clinical domains and regional datasets.

Results

After full processing, ~99% of vaccination records were retained as valid and standardized within the OMOP CDM. This high retention rate highlights both the quality of structured vaccination data in primary care and the robustness of the ETL pipeline.



The resulting dataset represents one of the largest standardized primary care vaccination databases in Belgium, with full compatibility with OMOP vocabularies and OHDSI analytical tools. The integration of LLM-assisted recovery reduced information loss without introducing significant noise, improving overall dataset completeness. The dataset enables longitudinal analyses and supports large-scale studies on vaccination coverage, adherence, and public health strategies.

Conclusion : This study demonstrates the feasibility of integrating large-scale primary care vaccination data into the OMOP CDM with minimal data loss and high coding validity. It highlights the maturity of vaccination data in GP systems and the value of combining rule-based and AI-assisted approaches. The resulting database provides a robust foundation for observational research and strengthens Belgium's contribution to international OMOP and OHDSI research networks.

OHDSI

M. **Borshchivska**, M. **Bastin,** T. **Helleputte**, A. **Kanfoud**, G. **Vanhalst** , R. **Verschuren**, O. **Latignies**, T. **Klein,** F. **Daue**, I. **Pollet**, S. **Arena** A. **Vandenberghe**