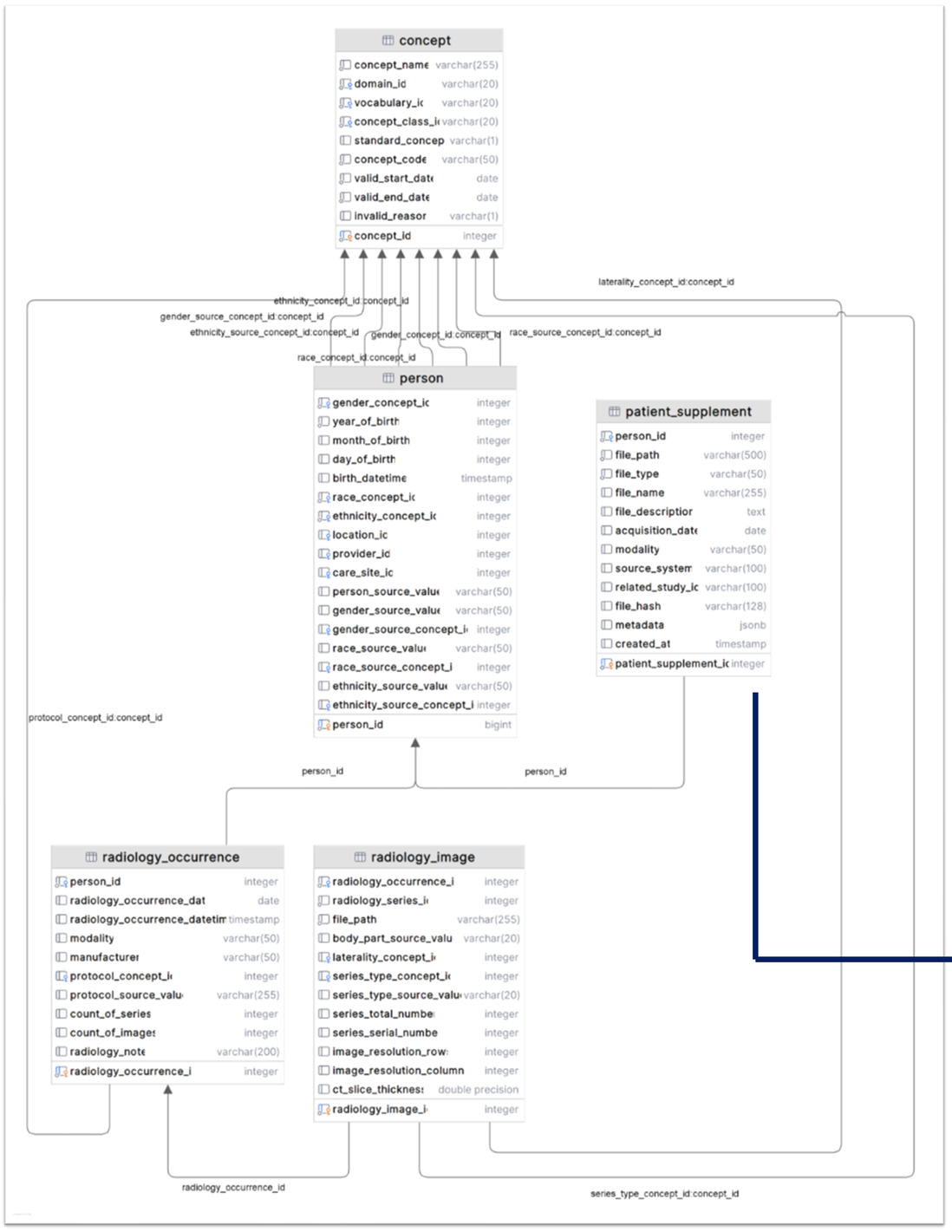
Extending OMOP with Object Storage for Cohort-Driven Access to Unstructured Data

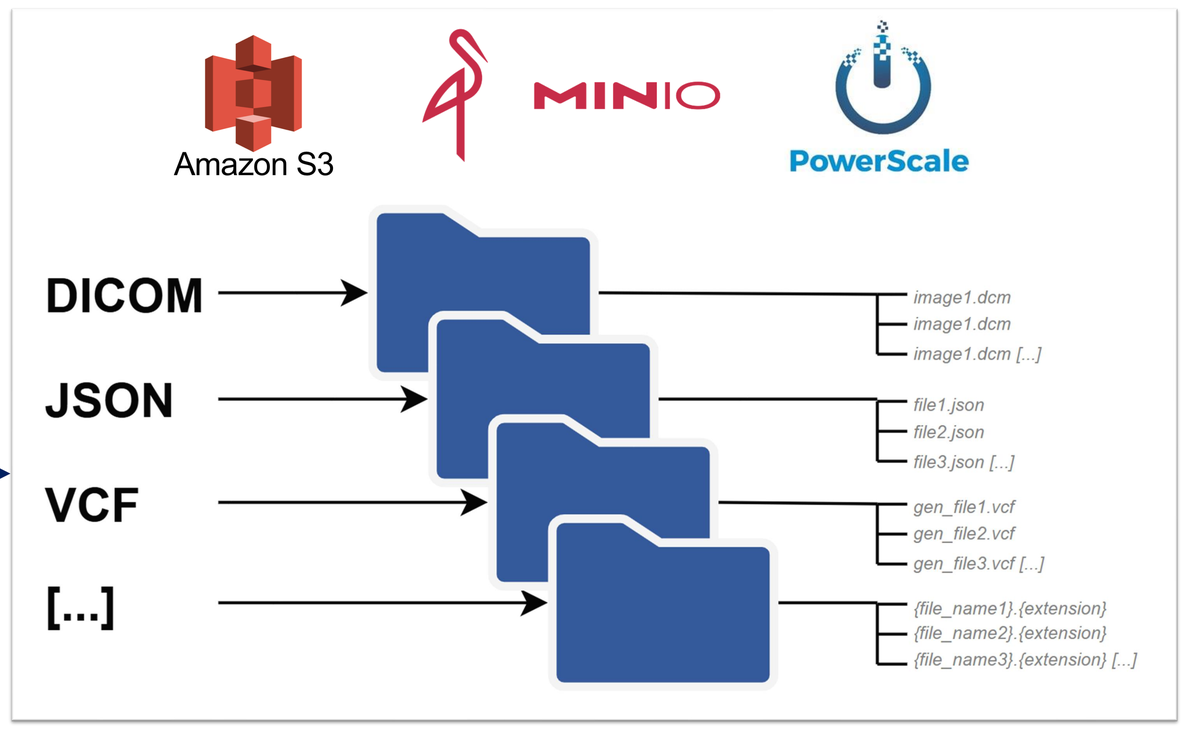
Title: ***Bridging OMOP patient records and patient supplement files through object storage solutions***

Funded by the European Union

The OHDSI OMOP Common Data Model (CDM) is widely used for harmonizing clinical data from Electronic Health Records (EHR) but provides limited native support for linking large, unstructured patient-level assets such as medical imaging, genomics, and documents. These data types are increasingly essential, particularly for the evaluation of AI as a medical device, where reproducible and cohort-driven access to supplementary files is critical. By design, the OMOP CDM cannot capture the full complexity or volume of such source data, making external storage and linkage unavoidable. Moreover, many analytical pipelines and medical software tools require direct access to the original source files rather than derived or summarized representations, reinforcing the need for robust mechanisms to associate OMOP records with their corresponding raw data assets. Several approaches have been proposed to address this limitation, such as domain-specific extensions including the Radiology Common Data Model (R-CDM) and more recently the Medical Imaging CDM (MI-CDM). However, these solutions are typically focused on a single data modality and lack the generality required to support heterogeneous unstructured data types. Any solution for integrating such data with OMOP must therefore be highly flexible, and deployable across diverse technical environments, while preserving the integrity and interoperability of the OMOP CDM.



This work was heavily inspired by the R-CDM, which links OMOP records to DICOM objects via external file references and structured metadata. We generalize this concept by introducing a lightweight extension table (***patient_supplement***) that links OMOP person_ids to object-stored files across multiple modalities. The approach enables nearly fully automated, scalable retrieval of patient supplement files for cohort-based analyses. It supports multiple storage backends, and large file volumes without impacting OMOP performance. The proposed pattern has been validated using local file systems as well as S3-compatible object storage solutions (MinIO, PowerScale).



This work demonstrates a pragmatic and OMOP-compatible extension pattern for integrating object storage with cohort-driven workflows without altering or constraining the core CDM. By layering a single, lightweight extension table on top of an existing OMOP instance, the approach preserves the OMOP CDM spirit while enabling access to supplementary patient-level data. The required setup is intentionally simple, relying on minimal schema additions and straightforward ETL logic for integrating object storage with OMOP without modifying the core CDM.

REALM
EU Project n°101095435

Frédéric Jung[1], Stef Rommes[1], Gökhan Ertaylan[1]

1. VITO, Vlaamse Instelling voor Technologisch Onderzoek, Mol, Belgium