

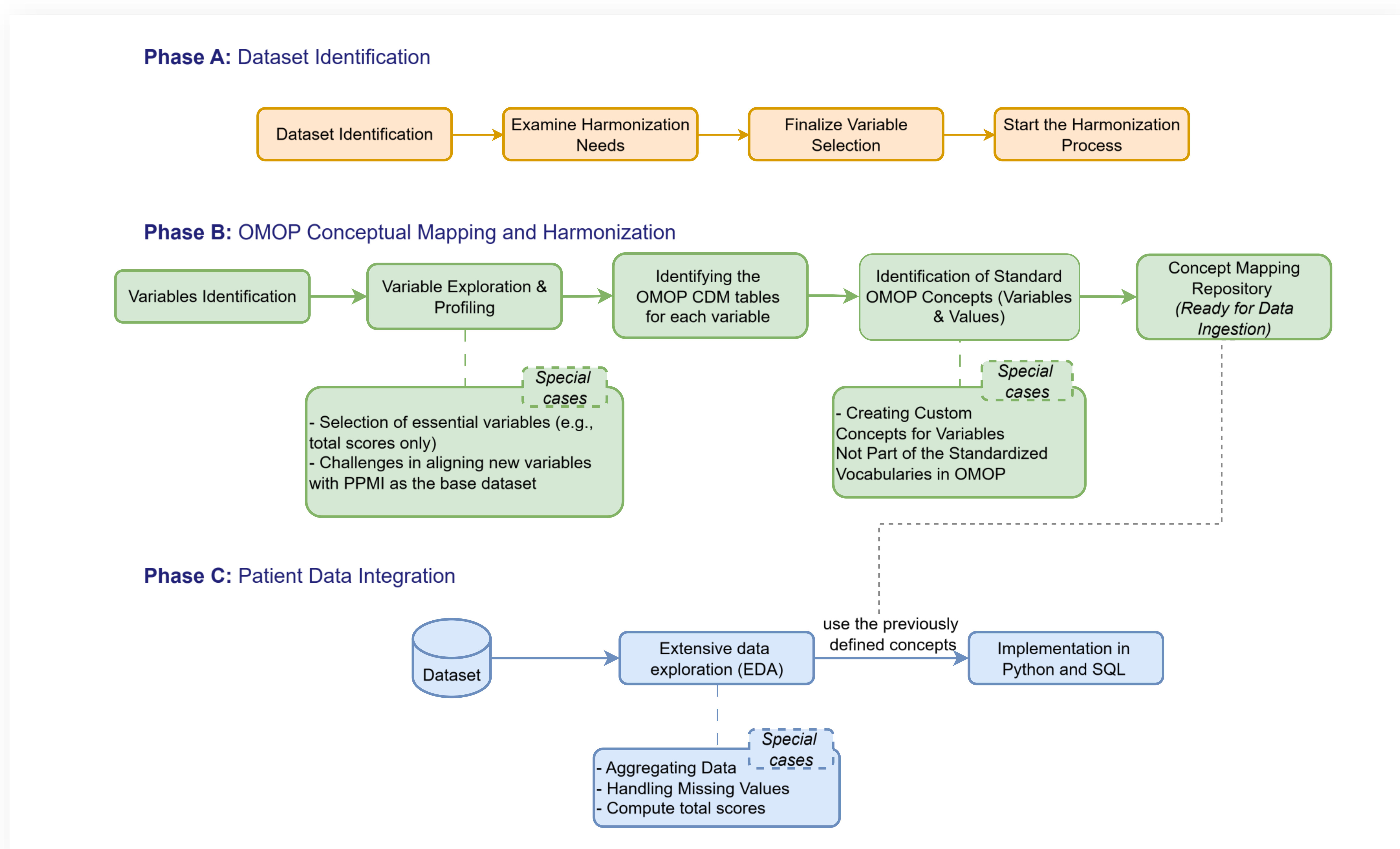
A systematic methodology for harmonization of Parkinson's disease research datasets to the OMOP CDM for interoperable and AI-ready analysis

Title: Disease-aware Harmonization of Parkinson's Disease Data to the OMOP Common Data Model

Background: Parkinson's disease research increasingly relies on integrating heterogeneous longitudinal datasets from multiple clinical studies. However, differences in data structure, terminology, and study design limit interoperability and cross-study analysis. The **OMOP Common Data Model (OMOP CDM)** provides a standardized framework for harmonizing observational healthcare data using controlled vocabularies such as **SNOMED CT, LOINC, and RxNorm**.

Methods

In the context of the AI-PROGNOSIS Horizon Europe project, we developed a **three-phase harmonization methodology** for transforming Parkinson's disease datasets into OMOP CDM. The workflow was developed using the **Parkinson's Progression Markers Initiative (PPMI)** dataset and it could be extended and applied to any PD-related dataset.



Results

For the purpose of the AI-PROGNOSIS project, we harmonized 2 datasets from retrospective and prospective studies with the proposed workflow. During the process we introduced **new vocabularies covering aspects of Parkinson's disease data not represented by existing standard concepts**. These vocabularies include concepts for some widely used questionnaires, such as MDS-UPDRS, RSBDSQ and HVLIT. The datasets cumulatively include > 5000 patients, 700+ free-text conditions and 800+ medication logs. We also created 182 new custom concepts required to capture the intricacies of PD-related data. We plan to contribute these new vocabularies to the OHDSI standards.

Limitation: The methodology was validated using 2 datasets that follow a research-study structure rather than routine clinical workflows. Future work will extend the approach to **additional Parkinson's disease** cohorts and real-world clinical datasets.

