# Overview covering the work since Mid March 2023 - 5th May 2023

**ETL Team:** Amelia Taylor, Gonjetso Chinyama, Moses Gwaza

Document written by: Gonjetso Chinyama  
Reviewed by: Amelia Taylor

1. Learning Pentaho
   * Using tutorials, we began the process with learning how Pentaho works and how we can use it to transform data.
   * We focussed on the use of Spoon and Kettle.
2. Running the WHO synthetic data through Pentaho
   * Using the ETLs from <https://github.com/tathagatabhattacharjee/Generic-IDSR-COVID-19-data-to-OMOP-6.0-under-INSPIRE-Project>, we ran the WHO synthetic data through Pentaho, taking note of how the data is being transformed and saved in the database.
   * We also learnt what changes are being made to the data from OMOP 6.0 to OMOP 5.4.
3. Mapping fields in Malawi IDSR synthetic data and WHO synthetic data
   * The Malawi IDSR does not have the same structure of the WHO synthetic data. Malawi IDSR has 51 fields whereas WHO data had 35. There are other differences between these datasets some of which will be covered later in this report.
   * We mapped out fields holding the same data in the Malawi IDSR and WHO data, and their data types.
   * There are several fields that exist in the Malawi data that do not exist in the WHO data. This will require additional mapping into OMOP.
4. Mapping the concept IDs in the WHO synthetic data and understand what they are for
   * The original ETLs we used had many concept IDs which we needed to understand before we ran the Malawi data. This helped us identify if there were any changes to be made.
   * There is use of constant concepts (we give a couple of examples in a separate document), e.g., for ethnicity, type of visit.
   * We also have different data elements in the Malawi IDSR and these needed to be mapped to OMOP concept, for example, names of vaccines or symptoms.
5. Identifying steps in the transformation that need modification to enable Malawi IDSR to run.
   * We modified the transformations to add vaccine name and ID for ‘Johnson & Johnson’
   * Added ‘Port of Entry’ path on patient types. The original transformations only had outpatient and inpatient.
   * In the travel destination step, changed ‘Outside Wakanda’ to ‘International’, and ‘Within Wakanda’ to ‘Indigenous’.
6. Running the Malawi IDSR data through Pentaho
   * We ran the data in Pentaho.
   * Throughout the process, we faced many issues which required us to resolve and run the transformations again as we worked through the errors. As we fixed the errors, it gave us a deeper understanding of how the transformations we designed as well as how Pentaho works overall.
   * We have documented the changes made and all the mapping documents separately.
7. Generating SQL script
   * + This script creates new columns from existing columns and changes other column names in the Malawi IDSR synthetic data to match those in the WHO synthetic data
     + We did this to allow us to run the Malawi data through the ETL without having to rename fields in all the transformations.

## Learning Pentaho

Learning Pentaho was the first step we carried out. We started by gathering resources and materials that could provide us with a comprehensive understanding of Pentaho's features and capabilities. We explored various [tutorials and guides](https://drive.google.com/drive/folders/1R0vjKySjlePUwpF3-ipZRv6z2lfOwmlN?usp=share_link) that helped us learn how Pentaho works and how it can be used to transform data.

Our learning process involved studying the basic concepts of Pentaho, such as its data integration, data mining, and business intelligence features. We also learned about the different tools and interfaces that Pentaho provides, such as the Pentaho Data Integration (PDI) tool, which is a powerful ETL (Extract, Transform, Load) tool that enables us to extract data from various sources, transform it into the desired format, and load it into a target destination.

We also gained an understanding of the Pentaho Data Integration (PDI) platform, which provides a user-friendly interface for designing and executing ETL jobs. We explored the different types of data sources and targets that Pentaho supports, including flat files, and databases.

Our learning process involved working through examples and exercises that helped us gain practical experience with Pentaho's features. This included creating data transformations to manipulate data and performing data analyses to gain insights from the transformed data. Through this process, we gained a deep understanding of Pentaho's capabilities and were able to apply this knowledge to our data transformation project.

## Running the WHO synthetic data through Pentaho

After we had familiarized ourselves with Pentaho, we moved on to the next stage of our project, which involved running the WHO synthetic data through Pentaho. To do this, we utilized the ETL (Extract, Transform, Load) files created by Tathagata Bhattacharjee from a publicly available [repository on GitHub](https://github.com/tathagatabhattacharjee/Generic-IDSR-COVID-19-data-to-OMOP-6.0-under-INSPIRE-Project). These ETLs were designed to transform the WHO synthetic data into the OMOP 6.0 format, which is a widely used data model for observational health data.

As we ran the WHO synthetic data through Pentaho using the ETLs, we made careful note of how the data was being transformed and saved in the database. This allowed us to gain a detailed understanding of the data transformation process and how the data was being modified to conform to the OMOP 6.0 data model.

We also took the opportunity to learn about the differences between OMOP 6.0 and its predecessor, OMOP 5.4. We discovered that OMOP 6.0 has several updates and enhancements that improve the quality and usability of the data model. For example, OMOP 6.0 includes new vocabulary concepts, improved mapping to standard terminologies, and expanded support for patient-level data.

By running the WHO synthetic data through Pentaho and examining the transformations made to the data, we gained valuable experience in using Pentaho to transform and integrate data from different sources. We also gained an understanding of how the OMOP data model works and the benefits it provides for analyzing and sharing health data. This knowledge was critical for the next stages of our project, where we focused on transforming and integrating the Malawi IDSR synthetic data.

## Mapping fields in Malawi IDSR synthetic data and WHO synthetic data

Once we had successfully run the WHO synthetic data through Pentaho and understood the transformations being made, we turned our attention to the next phase of our project, which involved mapping the fields in the Malawi IDSR synthetic data to the WHO synthetic data.

To begin this process, we first identified fields in the Malawi IDSR synthetic data that held similar data to fields in the WHO synthetic data. We then mapped these fields to each other and recorded their data types. This process allowed us to understand the similarities and differences between the two datasets and lay the groundwork for data integration.

During this mapping process, we also discovered that there were many fields in the Malawi data that did not exist in the WHO data. This was not unexpected, as the Malawi IDSR data contains information specific to Malawi's healthcare system that is not present in the WHO data. However, this presented a challenge for integrating the two datasets, as we needed to find ways to account for the missing data and ensure that the integrated dataset was complete and accurate.

## Concept IDs in the WHO synthetic data

We encountered numerous concept IDs in the WHO synthetic data that were derived from various medical dictionaries such as SNOMED and LOINC. Before proceeding with running the Malawi data through the ETL process, it was crucial for us to map these concept IDs and gain a comprehensive understanding of their meanings and purposes.

To accomplish this, we delved into the original ETLs that we had utilized for the WHO data transformation. These ETLs had incorporated concept IDs from different medical dictionaries to ensure standardized and consistent representation of medical concepts within the dataset. By examining these concept IDs in detail, we aimed to familiarize ourselves with the underlying terminologies and their associated definitions.

This mapping process served multiple purposes. First, it allowed us to gain a clear comprehension of the specific medical concepts represented by each concept ID. This knowledge was essential for accurately interpreting and analyzing the data within the WHO synthetic dataset. Second, it provided us with the opportunity to identify any potential changes or adaptations required for the concept mapping when integrating the Malawi IDSR synthetic data.

By thoroughly understanding the concept IDs and their significance within the WHO synthetic data, we were better equipped to assess the compatibility and applicability of these concepts to the Malawi IDSR dataset. This analysis enabled us to identify any variations or modifications that might be necessary to align the concept IDs and ensure the smooth integration of the two datasets.

## Identifying steps in the transformation that need modification to enable Malawi IDSR to run

As we embarked on the integration of the Malawi IDSR synthetic data into the data transformation pipeline, we identified several crucial steps within the existing transformations that required modification to accommodate the specific requirements of the Malawi data. By addressing these modifications, we ensured that the Malawi IDSR data could be seamlessly processed and integrated with the rest of the dataset. The key modifications made were as follows:

1. Addition of vaccine name and ID for 'Johnson & Johnson':

Within the transformations, we recognized the need to include specific information related to the 'Johnson & Johnson' vaccine. To accomplish this, we modified the **Modified JavaScript value 2** step within **/07 Vaccination/001 populate drug exposure.ktr** to incorporate the additional vaccine name and its corresponding ID. This modification allowed for accurate representation and analysis of 'Johnson & Johnson' vaccine-related data within the integrated dataset.

1. Introduction of 'Port of Entry' path on patient types:

The initial transformations solely classified patient types as either 'outpatient' or 'inpatient'. However, in the context of the Malawi IDSR data, it was essential to account for a distinct category related to 'Port of Entry'. To address this requirement, we modified the transformations (**/01 Demographics/006 populate visit detail.ktr**) to incorporate the 'Port of Entry' path within the patient type classification. This modification enabled appropriate categorization and analysis of data related to individuals entering through specific ports of entry.

1. Change of travel destination terminology:

To align the terminologies used in the travel destination step with the specifics of the Malawi IDSR data, we undertook necessary adjustments. Specifically, we modified the transformation process (**/06 Travel History/001 populate observation.ktr**) to change the label 'Outside Wakanda' to 'International' and 'Within Wakanda' to 'Indigenous'. This modification improved the clarity and consistency of travel destination categorization within the integrated dataset, enabling accurate analysis and interpretation of travel-related information.

By identifying and implementing these modifications, we ensured that the data transformation process accommodated the unique characteristics and requirements of the Malawi IDSR synthetic data.

## Running the Malawi IDSR data through Pentaho

After completing the necessary modifications to the transformations, we proceeded to run the Malawi IDSR data through Pentaho. This step involved executing the updated transformations on the Malawi IDSR dataset to facilitate data integration and achieve the desired outcome.

However, as with any complex data transformation process, we encountered several challenges along the way. These challenges necessitated problem-solving and troubleshooting efforts to address the issues that arose during the execution of the transformations. We worked diligently to resolve these problems, often requiring us to re-run the transformations multiple times. Each iteration and error resolution provided us with valuable insights into the design of the transformations and deepened our understanding of how Pentaho operates as a whole.

The process of identifying and resolving errors enhanced our knowledge of the intricate details of the transformations we had developed. It also allowed us to gain a more comprehensive understanding of Pentaho's functionalities and capabilities. Through hands-on experience and troubleshooting, we became proficient in leveraging Pentaho's features to ensure the successful execution of the data transformations.

Throughout this entire journey, we recognized the importance of documentation. As a result, we diligently documented all the changes made to the transformations, as well as maintaining separate mapping documents. This documentation serves as a valuable resource for future reference, enabling us to reproduce and replicate the project if needed. Additionally, it provides a comprehensive record of the modifications made during the integration of the Malawi IDSR data, aiding in understanding the steps taken and decisions made throughout the process.

In conclusion, running the Malawi IDSR data through Pentaho was a significant milestone in our project. Despite encountering various challenges, we successfully executed the transformations, resolved errors, and deepened our understanding of both the transformations and the Pentaho platform. The documentation of changes and mapping documents ensures the traceability of our work and provides a valuable resource for future analysis and replication.

### Challenges

**Mapping Complex Data Concepts:**

Mapping fields and concepts between different datasets, particularly when dealing with complex data structures, posed a significant challenge. The Malawi IDSR synthetic data and the WHO synthetic data employed different data models, terminologies, or data formats. Aligning and mapping these structures accurately required in-depth understanding of the data sources and meticulous analysis to ensure the correct integration of the data.

**Transformation Complexity:**

The complexity of the data transformation process presented its own set of challenges. The transformations originally developed for the WHO dataset involved a variety of operations, including data extraction, data cleansing, data enrichment, and data aggregation.

## Generate SQL script

Once we had mapped the fields in the Malawi IDSR synthetic data to the WHO synthetic data, we needed to prepare the Malawi data for integration with the WHO data. To do this, we wrote a SQL script that created new columns from existing columns in the Malawi IDSR synthetic data and changed the names of some columns to match those in the WHO data.

The purpose of this script was to align the data in the Malawi dataset with the data in the WHO dataset, making it possible to run the Malawi data through the same ETLs we used for the WHO data. By creating new columns and renaming existing columns, we could ensure that the Malawi data had the same structure as the WHO data, which simplified the integration process.

One advantage of this approach was that it allowed us to avoid having to rename fields in all the transformations we had created for the WHO data. This saved us a considerable amount of time and effort, as we were able to reuse most of the ETLs we had developed for the WHO data with minimal modification.

—------------- END OF DOCUMENT —-------------------------