Tools for predictive analysis of COVID-19 data in Malawi

Amelia V. Taylor
Inspire PEACH, Malawi University of Business and Applied Sciences *
ataylor@mubas.ac.mw

Abstract

We present an application of machine learning for preparing Malawi COVID-19 data for predictive analysis and patient characterisation studies using the OMOP common data model. We discuss the use of synthetic data for dta transformations and the use of clustering for cleaning and standardisation.

1 The Case of using OMOP CDM for COVID-19 Data in Malawi

The global impact of COVID-19 has spurred collaboration for data sharing, analysis, and technology dissemination. These collaborative efforts necessitated the integration of diverse datasets and the development of shared tools, definitions, and quality indicators. Communities such as the Observational Health Data Sciences and Informatics (OHDSI) [5] have expanded their tools to conduct COVID-19 studies. Originally established in the US in 2014 as the Observational Medical Outcomes Partnership, OMOP has evolved into a global consortium of academic and medical institutions. They collectively contribute to a comprehensive set of tools, including data models, datasets, experimental protocols, and database evaluation tools, all available in the public domain [3].

Collaborative studies involving disparate data sources transformed into the OMOP Common Data Model have been conducted and published. For instance, patient characterization studies spanned three continents: America, Europe, and Asia [2], while individual risk prediction for COVID-19 was explored using de-identified electronic health records [4]. However, in Malawi, characterization and prediction studies of COVID-19 patients exist only in small cohorts [1] and not on a nationwide scale, as the country lacks an integrated national EHR system. Although nationwide COVID-19 data was collected, the focus was primarily on reporting incidence numbers such as daily cases, deaths, and hospitalizations. Collecting patient-level data on a national level proved complex, resource-intensive, and did not yield analysis-ready data within the required timeframe.

We want to understand the extent to which Malawi can benefit from the OHDSI framework for leveraging their COVID-19 data. In this article, we present our approach to preparing the Malawi COVID-19 data for predictive analysis with OHDSI.

2 The Malawi COVID-19 Data

Malawi collected COVID-19 data using the Integrated Disease Surveillance and Reporting (IDSR) system. Surveillance for COVID-19 was meant to capture the four types of cases: suspected cases, confirmed cases, probable cases and contact case. Although initially several forms existed to collect data at various points and stages, e.g., ports of entry such as airports and land borders or at clinical facilities, one form emerged as the common data collection tool for COVID-19 surveillance. This was called the Case Based Surveillance and Reporting Form (CBSR form). It contained 51 variables

*This work is part of the project https://inspiredata.network and a collaboration with the Public Health Institute of Malawi.

5th Deep Learning Indaba Conference (DLI 2023).

grouped into nine sections or categories: (a) related to facility and clinical inclusion criteria; (b) demographics for the case; (c) vaccination information; (d) contact tracing information; (e) case symptoms history; (f) underlying conditions; (g) specimen information; (h) lab and test results information; (i) clinical outcome. The CBSR form served the basis of an national excel-based database containing all positive cases (people who had a positive COVID-19 test result). The line list database contained additional information such as hospitalisation and treatment.

Nationally, the Ministry of Health in Malawi through the Public Health Institute of Malawi published regular incidence statistics to the wider population. Aggregate and individual COVID-19 data was supposed to be entered into the national One Health Surveillance Platform (OHSP), which was setup for that purpose. However, this was only partly done and the most reliable and comprehensive database remain in an Excel based form of the "COVID-19 line list".

Throughout our project, we encountered significant delays in acquiring actual data, despite obtaining the necessary ethical clearances early on. However, when we finally obtained the data, extensive cleaning and preparation were required due to its specific characteristics:

- Data Collection and Sources: he data was obtained by merging and copying information from numerous smaller Excel files originating from different reporting districts in Malawi.
- Lack of Uniform Coding: One notable challenge was the absence of uniform coding, even for fields where national codes were available. For instance, facility codes and district codes did not follow a consistent coding system, leading to inconsistencies and difficulties in data management.
- Presence of Large Gaps: The data contained significant gaps, indicating missing information for certain records or fields. These gaps posed challenges in ensuring comprehensive and complete data analysis.
- Incomplete Fields: Multiple fields within the dataset were not adequately filled with the expected information, resulting in incomplete data for analysis and interpretation.

Given these characteristics, substantial effort was dedicated to cleaning and preparing the data to address inconsistencies, missing values, and incomplete fields. These data cleaning activities were necessary to ensure the reliability and accuracy of subsequent analyses and findings. Therefore, to prepare this data for analysis a large amount of data cleaning and standardisation needs to take place. The size of the data was over 74,587 records and spanned the years 2020 to 2022.

2.1 Using Synthetic Data to setup the OMOP pipeline

To avoid delays in the development of the pipeline to OMOP, and in anticipation of obtaining real data, we generated synthetic data that follows the profile of the collection instrument, the CBSR form. Our approach was novel and opens opportunities to generate synthetic datasets to facilitate research [2].

We used Python libraries to generate values for the 51 variables. To make the data as realistic as possible, given that we do not have access to a dataset of real data, we set up a database of categories (options) to be used for some of the variables. The data we generated satisfies the following individual level data fidelity metrics:

- Names and Gender Correspondence: for example typical male names correspond to Male (sex). We generated only first names and codes for Surnames to avoid a situation where the surnames have been all European.
- Facilities and Districts are linked: a facility is expected to be located in a certain district.
- Linked and Staggered Dates: We implemented a realistic order/priority of dates. For example, the date when the specimen was sent to the lab cannot be after the date in which the result is sent back to the health center.

The availability of the synthetic data allowed us to progress with mapping concepts to vocabularies and develop the ETLs.

[2] `https://github.com/Inspire-Mubas/Malawi-IDSR-COVID-19-Synthetic-DataSet`

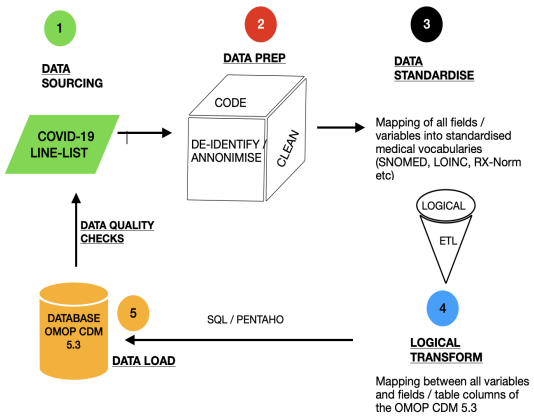

Figure 1: Data Preparation Pipeline to OMOP CDM

3 The Data Preparation Pipeline

The data preparation pipeline starts with the clean source data (Malawi COVID-19 Line Lists) and includes 4 steps: (I) mapping of health/medical terms into standardised vocabularies; (II) writing the logical transformations of the data from line-lists to the tables structure of the OMOP CDM 5.4; (III) writing scripts to load data into a physical database and (IV) running data quality checks on the transformed data to ascertain the correctness of the transformations. We used Pentaho and SQL to load data into a PostgreSQL database containing the tables of the CDM 5.3 model. Full code and details are found in our github repository [3]. Pentaho Data Integration (PDI) enables the definition of workflows and jobs for transforming data from a source to a target format. Each step represents a specific group of data elements, such as demographics and disease condition. This ensures a systematic transformation and loading of the source data into the target OMOP Common Data Model tables, following a predefined order.

3.1 Clustering and Text Similarity for Data Preparation

After sourcing real data, the first task consisted of cleaning, coding and de-identifying. We utilized text similarity and clustering methods to perform data cleaning on several categorical variables in the COVID-19 Line List. For this we used an open-source tool called OpenRefine [4]. We manually created, reviewed, and obtained clusters for variables that required extensive cleaning, such as facility names, occupation names, and specimen types. From these clusters we generated training datasets that will be used to develop a model capable of automatically cleaning, labeling, and coding the data.

To illustrate the significance of this process, let's consider the field of **occupation**. During the COVID-19 pandemic, it was important to stratify occupations based on their risk of exposure to the disease for public health analysis. Certain occupations were identified as higher risk. Therefore, having clean and well-categorized occupation values in our dataset is vital for applying standardisation and for conducting predictive analyses.

As for other fields, there were large variations in names of occupations, some mentioned the name of a workplace or organization some mentioned a job function. Therefore, we applied several methods for detecting clusters: key collision algorithms, phonetic fingerprinting and distance based algorithms such as nearest neighbour. Phonetic clustering was useful because there are many cases in our data, where occupations were spelled incorrectly, or they were spelt as they sounded to the recorder.

[3] https://github.com/Inspire-Mubas/Pentaho-ETL-Malawi-IDSR
[4] https://openrefine.org/

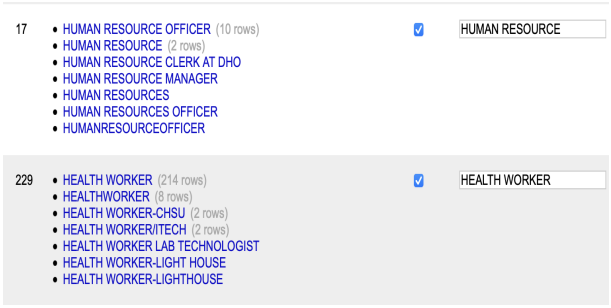

Figure 2: Examples of obtaining clusters for the training set.

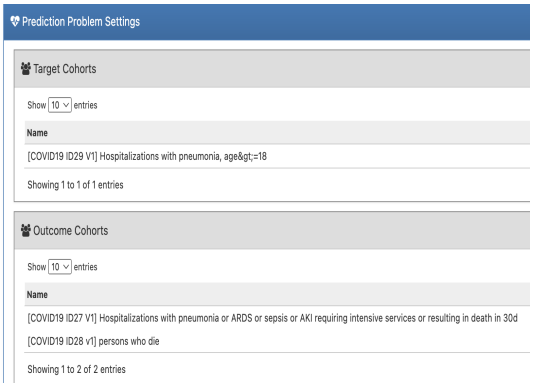

Figure 3: Setting up prediction studies with Atlas in OMOP CDM

3.2 Data Standardisation

Data standardisation is the process of encoding variables and values in the source data with concepts defined by medical vocabularies. For example:

- A laboratory test that is *positive for COVID-19* is mapped to **4126681** from the Condition domain of the SNOMED vocabulary.,
- The condition *COVID-19* is mapped to **37311061** from the measurement / diagnosis domains of the SNOMED vocabulary.

Data standardisation and the use of the OMOP CDM can facilitate the definition of phenotype (i.e. patients having specific conditions). Patients with respiratory co-morbidities or patients with cardiac co-morbidities are examples of phenotype.

4 Conclusion

We are presenting the steps needed to prepare the Malawi COVID-19 data for predictive analysis into OMOP. Our data cleaning approach involved manual cleaning, and the use of cluster analysis for creating training datasets. Data standardisation and ETL used synthetically generated datasets. We also present the use of OMOP and OHDSI tools for standardising and enabling characterisation and predictive studies.

Acknowledgments and Disclosure of Funding

This research was part of the INSPIRE PEACH project which received funding from IDRC under AI4COVID.

References

[1] Catherine Anscombe, Samantha Lissauer, Herbert Thole, Jamie Rylance, Dingase Dula, Mavis Menyere, Belson Kutambe, Charlotte van der Veer, Tamara Phiri, Ndaziona P. Banda, Kwazizira S. Mndolo, Kelvin Mponda, Chimota Phiri, Jane Mallewa, Mulinda Nyirenda, Grace Katha, Henry Mwandumba, Stephen B. Gordon, Kondwani C. Jambo, Jennifer Cornick, Nicholas Feasey, Kayla G. Barnes, Ben Morton, Philip M. Ashton, Wezzie Kalua, Peter Mandala, Barbara Katutula, Rosaleen Ng'oma, Steven Lanken, Jacob Phulusa, Mercy Mkandawire, Sylvester Kaimba, Sharon Nthala, Edna Nsomba, Lucy Keyala, Beatrice Chinoko, Markus Gmeiner, Vella Kaudzu, Bridget Freyne, Todd D. Swarthout, Pui Ying Iroh Tam, Simon Sichone, Ajisa Ahmadu, Grace Stima, Mazuba Masina, Oscar Kanjewa, Vita Nyasulu, End Chinyama, Allan Zuza, Brigitte Denis, Evance Storey, Nedson Bondera, Danford Matchado, Adams Chande, Arthur Chingota, Chimenya Ntwea, Langford Mkandawire, Chimwemwe Mhango, Agness Lakudzala, Mphatso Chaponda, Percy Mwenechanya, Leonard Mvaya, Dumizulu Tembo, Marc Y.R. Henrion, James Chirombo, Paul Kambiya, Clemens Masesa, and Joel Gondwe. A comparison of four epidemic waves of covid-19 in malawi; an observational cohort study. *BMC Infectious Diseases*, 23, 2023.

[2] Edward Burn, Seng Chan You, Anthony Sena, Kristin Kostka, Hamed Abedtash, Maria Tereza F Abrahao, Amanda Alberga, Heba Alghoul, Osaid Alser, Thamir M Alshammari, Carlos Areia, Juan M Banda, Jaehyeong Cho, Aedin C Culhane, Alexander Davydov, Frank J DeFalco, Talita Duarte-Salles, Scott L DuVall, Thomas Falconer, Weihua Gao, Asieh Golozar, Jill Hardin, George Hripcsak, Vojtech Huser, Hokyun Jeon, Yonghua Jing, Chi Young Jung, Benjamin Skov Kaas-Hansen, Denys Kaduk, Seamus Kent, Yeesuk Kim, Spyros Kolovos, Jennifer Lane, Hyejin Lee, Kristine E Lynch, Rupa Makadia, Michael E Matheny, Paras Mehta, Daniel R Morales, Karthik Natarajan, Fredrik Nyberg, Anna Ostropolets, Rae Woong Park, Jimyung Park, Jose D Posada, Albert Prats-Uribe, Gowtham A Rao, Christian Reich, Yeunsook Rho, Peter Rijnbeek, Selva Muthu Kumaran Sathappan, Lisa M Schilling, Martijn Schuemie, Nigam H Shah, Azza Shoaibi, Seokyoung Song, Matthew Spotnitz, Marc A Suchard, Joel Swerdel, David Vizcaya, Salvatore Volpe, Haini Wen, Andrew E Williams, Belay B Yimer, Lin Zhang, Oleg Zhuk, Daniel Prieto-Alhambra, and Patrick Ryan. An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. *medRxiv*, 2020.

[3] George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, Johan Van Der Lei, Nicole Pratt, G. Niklas Norén, Yu Chuan Li, Paul E. Stang, David Madigan, and Patrick B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In *Studies in Health Technology and Informatics*, volume 216, 2015.

[4] Tarun Karthik Kumar Mamidi, Thi K. Tran-Nguyen, Ryan L. Melvin, and Elizabeth A. Worthey. Development of An Individualized Risk Prediction Model for COVID-19 Using Electronic Health Record Data. *Frontiers in Big Data*, 4, 2021.

[5] Ines Reinecke, Michéle Zoch, Christian Reich, Martin Sedlmayr, and Franziska Bathelt. The usage of OHDSI OMOP a scoping review. In *Studies in Health Technology and Informatics*, volume 283, 2021.

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.