Mapping the population data of England into OMOP CDM

Leveraging big data technologies in a Secure Data Environment (SDE)

Mehrdad A. Mizani[1], Jadene Lewis[1], Rouven Priedon[1], Silvia Jimenez[2], Shirah Cashriel[2], Emma Gesquire[2], Anne Li[2], Thomas Bolton[1], John Nolan[1], Angela Wood[1]

[1]British Heart Foundation Data Science Centre, Health Data Research UK, London, UK
[2]edenceHealth NV

British Heart Foundation Data Science Centre
Led by Health Data Research UK

Background

Whole-population electronic health records (EHRs) from a range of sources are essential for advancing national health and healthcare by enabling comprehensive and accurate examination of health trends, disease patterns, treatment outcomes, and public health interventions. Moreso, they represent all ages, ethnicities, socioeconomic, and clinical characteristics, offering insights into health disparities and potential strategies for addressing them effectively. Motivated by the public health importance of fully understanding the health impact of the COVID-19 pandemic, the BHF Data Science Centre, led by Health Data Research (HDR) UK, in collaboration with the user community and public health organisations, has driven efforts in efficient, cost-effective and safe access to and analysis of whole-population EHRs through the CVD-COVID-UK/COVID-IMPACT research programme. The programme supports approved researchers to access and analyse nationally collated EHRs and health audits in the UK via national Secure Data Environments (SDEs). This includes whole-population datasets from England (~57 million patients), Scotland (~5.5 million), and Wales (~3.2 million)[1].

Analyses of these data have generated high-quality research findings that translate into real-world health improvements[2-5]. For example, our cardiovascular disease and COVID-19 research, conducted on behalf of the Chief Medical Officer to the UK Government, informed the Department of Health and Social Care secondary prevention policies[6]. However, the lack of standardisation in the data structure and clinical coding presents a substantial obstacle to the rapid utilisation of the datasets and integration across SDEs. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) offers a promising solution, enabling real-world studies on larger datasets, incorporating more patients and variables, utilising standardised tools, and facilitating federated analytics.

In this study, we present the progress of mapping primary care, hospital admissions, deaths, and National Heart Failure Audit (NHFA) data available via NHS England's (NHSE) SDE into the OMOP CDM v5.4. Data are available in a Databricks environment with Spark technologies to manage the large-scale longitudinal data. We conducted this project as a candidate data partner of the European Health Data and Evidence Network (EHDEN). The mapping pipeline was co-developed and tested in dummy and real datasets in collaboration with edenceHealth, an EHDEN-certified Small and Medium-sized Enterprise (SME). Due to the secure nature of the NHSE SDE and the large scale of the whole-population EHRs, we encountered several challenges. For example, the inability to provide edenceHealth with access to the SDE, due to data access permissions, meant that only data dictionaries were visible to the Extract-Transform-Load (ETL) designers. The absence of real-time Git operations further complicated collaborative development. Additionally, the lack of necessary software, such as Java for R-based Observational Health Data Sciences and Informatics (OHDSI) tools, and restrictions in installing non-CRAN R packages further complicated the process. Performance issues with OHDSI scripts on big data in the Spark environment added to the challenges.
However, through a collaborative team science approach and a multi-stage ETL development process, we overcame these challenges. This approach facilitated the successful mapping of datasets and provided valuable insights and best practices for handling large-scale real-world data.

Methods

In the first phase of the project, we mapped General Practice Extraction Service (GPES) Data for Pandemic Planning and Research (GDPPR), Hospital Episode Statistics Admitted Patient Care (HES APC), Civil Registration of Deaths, and NHFA datasets. GDPPR includes primary care records of patients alive on or born after 1 November 2019, excluding opt-outs[7]. Clinical events and prescribed medications are coded in SNOMED-CT*. HES APC contains hospital episodes, with clinical codes in ICD-10§ and procedures[8] in OPCS4‡. The Civil Registration of Deaths includes death records with causes of death coded in ICD-9 and ICD-10[9].

NHFA audits patients with unscheduled hospital admissions discharged with heart failure as the primary diagnosis[10], using a categorical coding that requires manual mapping to OMOP standard concepts. **Figure 1** illustrates the annual number of patients per dataset.

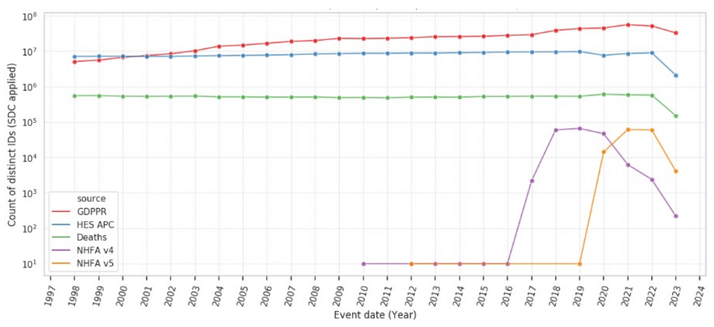

Figure 1. *Number of patients per dataset in logarithmic scale*

To prepare the structural mapping document, we created small dummy datasets from the data dictionary to facilitate the development of the ETL by edenceHealth. The ETL was developed in three phases: a) development and testing outside the SDE, b) testing in two stages within the SDE, and c) deploying to the production environment. The initial ETL design was performed by edenceHealth using the dummy dataset in a local Spark environment and Databricks. The code was manually transferred into the SDE, where an identical repository was created in the internal GitLab system. The ETL was first tested on a staging area with a sample of 5000 patients to verify its compatibility with the specifications of Databricks in the SDE. We employed a team science approach, leveraging agile methodology to streamline collaborative problem-solving. The ETL was applied to the entire dataset in the second testing phase to identify issues with large-scale data and performance bottlenecks, as shown in **Table 1**. Finally, the ETL was applied to the entire dataset, with the Delta tables saved for use in Achilles and Data Quality Dashboard (DQD) tools.

Results

Figure 2 shows the OMOP CDM tables completed in the first phase of the project. The location table contains the Lower Layer Super Output Areas (LSOA) of the general practitioner (GP) or the patient. The provider table includes the specialities of healthcare providers in HES APC. Visit occurrence contains the largest number of rows due to the high number of GP visits and HES APC admissions. Visit details correspond to individual episodes in HES APC. In the first phase of the project, the drug exposure table contains only prescribed medication. In the next phase of the project, we aim to add dispensed medications to the drug exposure table, include additional OMOP tables (e.g., device exposure), enhance the mapping of non-standard to standard concepts (e.g., in NHFA), and improve the ETL quality based on DQD outputs. As we progress towards completing the EHDEN project's final report, we are running Achilles and DQD queries on the OMOP tables and applying statistical disclosure controls for export by NHS England[11].

* SNOMED-CT: Systematized Medical Nomenclature for Medicine, Clinical Terminology
§ ICD: International Classification of Diseases
‡ OPCS: Office of Population Censuses and Surveys

Table 1. *Major issues of OMOP CDM ETL on real-world big data and the utilisation of OHDSI tools in NHSE SDE*

	Issues	Challenges	Solutions
Spark environment	Data skewness	- Data skewness due to relatively small lookup tables	+ Join-based mapping with broadcast
	Cluster crash or performance bottleneck	- The cluster is multi-tenant - Memory overhead with Python data structures - Sub-optimal PySpark User Defined Functions (UDFs)	+ Replacing Python data structure with PySpark data frames + Using built-in PySpark SQL functions
Scale of the data	Negative keys	- Maximum positive integer= 2,147,483,647	+ Big integer data type for primary/foreign keys with potentially large counts
	Non-unique or non-deterministic ID column	- Spark is not designed for globally unique IDs - Lack of primary key constraints in Spark - De-normalised tables and no unique row index - Non-unique, non-deterministic, non-consecutive monotonically increasing IDs	+ Creating a unique proxy column in using PySpark UUID combined with other columns + Employing an RDD-based (Resilient Distributed Datasets) indexing and window functions based on the unique proxy column
SDE restrictions	Single shared database/schema	- No admin rights - It is not possible to create different schemas (e.g., CDM, Write, Vocabularies)	+ No hard-coded schema/table names in the code + Parametrising schema and table names
	Naming conventions	- SDE naming convention guidelines (e.g., naming a table as "PERSON" is not recommended) - Multiple instances of OMOP CDM with standard table names is not possible in the same schema	
R in SDE	R tools do not run in the SDE	- Lack of local installation of Java in SDE - Non-CRAN R tools are not available - Read-only access from R environment to the database - R memory overhead	+ Using SQL-only mode + De-coupling database connection and query functionalities + De-constructing incompatible R functions and simulate them in PySpark
	Queries fail to run on large datasets	- Nested queries result in large Spark transformation plans - Some nested queries create small views (e.g., year and month combinations) leading to data skewness	+ Breaking down nested and run them with broadcast and persist functionalities + Translating the query into Spark-SQL or PySpark for debugging + Optimising the query with built-in PySpark SQL functions

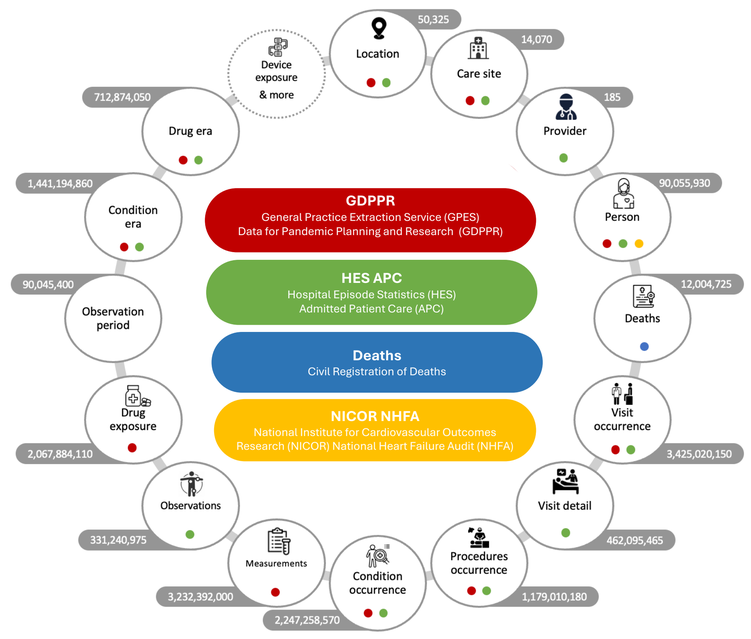

Figure 2. *Completed OMOP tables, and the number of rows per table, in the first phase of the project*

Conclusion

We have described our mapping of whole-population EHRs in England into the OMOP CDM v5.4. The vast scale of these datasets and the secure nature of the SDE challenged the conventional implementation of the OMOP CDM and the use of OHDSI tools and caused Spark performance bottlenecks. As SDEs become the standard for delivering health and social care data for research, in line with the NHS's "Data Saves Lives" strategy[12], and with Spark and Databricks environments becoming the default infrastructure, innovative approaches to OMOP ETL design and the utilisation of OHDSI tools are imperative.

Our project provides valuable insights into the practical implications of implementing an OMOP CDM project within a real-world, big-data environment in SDEs. It highlights the importance of optimising ETL and ensuring OHDSI tools and guidelines are compatible with SDEs. Moving forward, we aim to systematically assess ETL optimisation and SDE-compatible OHDSI tools while contributing to national and international federated studies leveraging the wealth of national data in SDEs.

References

1. CVD-COVID-UK/COVID-IMPACT TRE Dataset Provisioning Dashboard. United Kingdom (UK): BHF Data Science Centre [Updated 2024 June 13; cited 2024 July 1]; [about 1 p.]. Available from: https://bhfdatasciencecentre.org/wp-content/uploads/2024/06/240613-CVD-COVID-UK-COVID-IMPACT-TRE-Dataset-Provisioning-Dashboard.pdf
2. Ip S, North TL, Torabi F, Li Y, Abbasizanjani H, Akbari A, Horne E, Denholm R, Keene S, Denaxas S, Banerjee A. Cohort study of cardiovascular safety of different COVID-19 vaccination doses among 46 million English adults. medRxiv. 2024 Feb 13:2024-02.
3. Pineda-Moncusí M, Allery F, Delmestri A, Bolton T, Nolan J, Thygesen JH, Handy A, Banerjee A, Denaxas S, Tomlinson C, Denniston AK. Ethnicity data resource in population-wide health records: completeness, coverage and granularity of diversity. Scientific Data. 2024 Feb 22;11(1):221.
4. Kerr S, Bedston S, Cezard G, Sampri A, Murphy S, Bradley DT, Morrison K, Akbari A, Whiteley W, Sullivan C, Patterson L. Undervaccination and severe COVID-19 outcomes: meta-analysis of national cohort studies in England, Northern Ireland, Scotland, and Wales. The Lancet. 2024 Feb 10;403(10426):554-66.
5. Thygesen JH, Tomlinson C, Hollings S, Mizani MA, Handy A, Akbari A, Banerjee A, Cooper J, Lai AG, Li K, Mateen BA. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. The Lancet Digital Health. 2022 Jul 1;4(7):e542-57.
6. Knight R, Walker V, Ip S, Cooper JA, Bolton T, Keene S, Denholm R, Akbari A, Abbasizanjani H, Torabi F, Omigie E. Association of COVID-19 with major arterial and venous thrombotic diseases: a population-wide cohort study of 48 million adults in England and Wales. Circulation. 2022 Sep 20;146(12):892-906.
7. General Practice Extraction Service (GPES) Data for pandemic planning and research: a guide for analysts. United Kingdom (UK): NHS England [Updated 2024 June 25; cited 2024 July 1]; [about 12 p.]. Available from: https://digital.nhs.uk/coronavirus/gpes-data-for-pandemic-planning-and-research/guide-for-analysts-and-users-of-the-data
8. Hospital Episode Statistics. United Kingdom (UK): NHS England [Updated 2024 June 11; cited 2024 July 1]; [about 3 p.]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/hospital-episode-statistics
9. Civil Registration of Death. United Kingdom (UK): NHS England [Updated 2024 June 11; cited 2024 July 1]; [about 2 p.]. Available from: https://digital.nhs.uk/services/data-access-request-service-dars/dars-products-and-services/data-set-catalogue/civil-registrations-of-death
10. National Heart Failure Audit (NHFS). United Kingdom (UK): NICOR [Updated unknown; cited 2024 July 1]; [about 1 p.]. Available from: https://www.nicor.org.uk/national-cardiac-audit-programme/heart-failure-audit-nhfa
11. Output your results. United Kingdom (UK): NHS England [Updated 2024 April 12; cited 2024 July 1]; [about 4 p.]. Available from: https://digital.nhs.uk/services/secure-data-environment-service/log-in/user-guides/output-your-results
12. Secure data environment for NHS health and social care data – policy guidelines. United Kingdom (UK): Department of Health & Social Care [Updated 2022 December 23; cited 2024 July 1]; [about 8 p.]. Available from: https://www.gov.uk/government/publications/secure-data-environment-policy-guidelines/secure-data-environment-for-nhs-health-and-social-care-data-policy-guidelines

This work was carried out with the support of the BHF Data Science Centre, led by Health Data Research UK (BHF Grant no. SP/19/3/34678).

This project is supported by the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement no 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

This work was supported by Health Data Research UK which receives its funding from HDR UK (HDRUK2024.0336) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and Cancer Research UK.