OmopSketch: An R package to characterise your OMOP mapped database

Marta Alcalde-Herraiz[1], Yuchen Guo[1], Mike Du[1], Edward Burn[1], Martí Català[1]
[1]Medical Sciences Division, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences in the University of Oxford

Background

Most data used in observational healthcare research were not originally collected for research purposes. For instance, electronic health records (EHRs) are captured to document essential information for patient's care, while administrative claims are mainly collected to facilitate cost allocation to payers. This can lead to a potential mismatch between the quality and type of data collected, and the data requirements of specific research study. To ensure the validity and reliability of our research findings and conclusions, it is crucial to have a thorough understanding of the database we are analysing. This includes having a general overview of the available data over time, assessing its quality, and determining its suitability for our specific research study.

Currently there exist two commonly used OHDSI packages that cover part of these needs: Achilles and DQD. Achilles (Automated Characterisation of Health Information at Large-scale Longitudinal Evidence Systems)[1] provides an overall characterisation and visualisation of the database; and DQD (Data Quality Dashboard)[2] performs quality control across the different tables and columns to ensure they meet the given specifications.

However, these packages do not always provide enough information to determine if a database meets the necessary criteria to answer a specific clinical question. Metadata to characterise the follow-up period of the potential study participants, their observation period, and trends in the use of concepts over time, for example, would be extremely helpful to assess the suitability of a database before starting to conduct a study.

Here we present *OmopSketch,* an R package that characterises the OMOP tables contained in a database with the aim to further assess its suitability for specific research studies.

Methods

The *OmopSketch* package is written in R (version 4.3.2) and is built on top of *Omopgenerics*[3], *CDMConnector*[4], and *tidyverse*[5]. The package was tested using *testthat* framework against synthetic data generated with *omock*[6]. The package was tested with different database management systems (DBMS) such as duckDB, PostgreSQL, SqlServer and Databricks.

The package processes the OMOP tables contained in the database to summarise essential information that can be used to assess its suitability for a specific study. The package provides both basic functionality that is only expected needed to run once per database release but also more functionality that may be run on a study-by-study basis for further exploration of the domains and likely participants of a particular study.

OmopSketch focuses on key characteristics of the clinical tables, such as the number of records and subjects, the number of concepts mapped, the domain and type of these concepts, and source vocabularies. Additionally, it tracks trends in records over time. *OmopSketch* also provides insights into the database's observation period, including trends in individual observations, follow-up histories and person-days contribution.

Results

OmopSketch can be freely accessed via the public GitHub Repository (https://github.com/oxford-pharmacoepi/OmopSketch) and will be available on CRAN when a first stable release is ready.

To illustrate the main functionalities, consider the following use-case scenario: *incidence study for some drugs from 1960 to 1999 using GiBleed Eunomia data*. Before starting our research, and following a comprehensive database overview and quality checks, we need to ascertain if our database is suitable for such a study.

Firstly, we characterise the *drug_exposure* table. The code presented in Figure 1A uses *OmopSketch* to generate, first, a standardised result object[3], and afterwards, to create a *gt* table. The resultant *html* table can be seen in Figure 1B.

```
result <-
  summariseOmopTable(
    omopTable = cdm$drug_exposure,
    recordsPerPerson = c(
      "mean", "sd", "median",
      "q25", "q75"),
    inObservation = TRUE,
    standardConcept = TRUE,
    sourceVocabulary = FALSE,
    domainId = TRUE,
    typeConcept = TRUE)

tableOmopTable(result)
```

Variable	Level	Estimate	cdm_name GiBleed
drug_exposure			
Number of subjects	-	N (%)	2,694 (100.0%)
Number of records	-	N	67,707
Records per person	-	median [IQR]	25 [22 - 28]
		mean (sd)	25.13 (5.25)
In observation	No	N (%)	251 (0.4%)
	Yes	N (%)	67,456 (99.6%)
Standard concept	Standard	N (%)	67,707 (100.0%)
Source vocabulary	CVX	N (%)	25,710 (38.0%)
	NDC	N (%)	2,694 (4.0%)
	No matching concept	N (%)	35 (0.1%)
	RxNorm	N (%)	39,268 (58.0%)
Domain	Drug	N (%)	67,707 (100.0%)
Type concept id	Prescription written (38000177)	N (%)	41,997 (62.0%)
	Dispensed in Outpatient office (581452)	N (%)	25,710 (38.0%)

Figure 1. Summary of GiBleed drug exposure table. (A) Code to summarise and visualise an OMOP table (drug_exposure) with *OmopSketch*. **(B)** Visualisation (*gt* table) of the summary.

To further characterise the table, we explored the records trend to identify any unexpected abrupt disruptions. Figure 2A shows how to first create a summarised result containing the table's record count stratified by age group, and then visualising it. The resulting output is shown in Figure 2B.

The package also provides functionality to characterise the observation period to (1) explore the trend and (2) extract number of individuals in observation.

Documentation of all functionalities can be found in: https://oxford-pharmacoepi.github.io/OmopSketch/reference

```
result <- summariseRecordCounts(
  omopTable = cdm$drug_exposure,
  unit = "year",
  unitInterval = 1,
  ageGroup = list(
    "<65" = c(0, 64),
    ">=65" = c(65, 150)))

plotTableCounts(result)
```

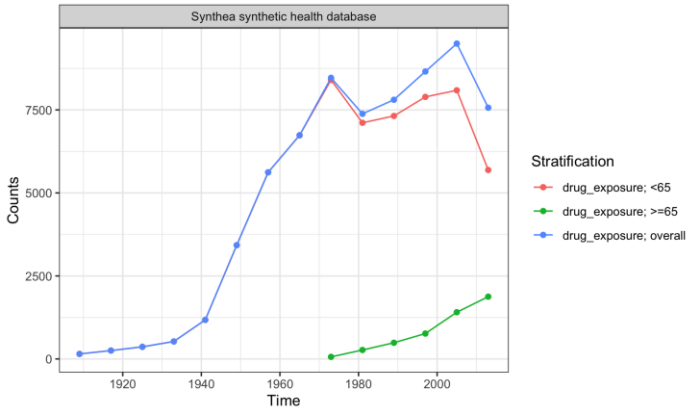

Figure 2. Record count for GiBleed drug exposure table. (A) Code to summarise and visualise the record count for an OMOP table (drug_exposure) with *OmopSketch*. **(B)** Visualisation (*ggplot2*) of the summary.

Conclusion

OmopSketch offers a variety of functions designed to contextualise your OMOP-mapped database and assist in evaluating the suitability for specific studies. The package provides insights into clinical table, observation period trends, and follow-up data. Additionally, we expect to incorporate further analysis exploring concepts counts and individual characteristics.

References

1. DeFalco F, Ryan P, Schuemie M, Huser V, Knoll C, Londhe A, Abdul-Basser T, Molinaro A (2023). *Achilles: Achilles Data Source Characterization*. R package version 1.7.2.
2. Blacketer C, Schuemie FJ, Ryan PB, Rijnbeek P (2021). "Increasing trust in real-world evidence through evaluation of observational data quality." *Journal of the American Medical Informatics Association*, **28**(10), 2251-2257. https://doi.org/10.1093/jamia/ocab132.
3. Catala M, Burn E (2024). *omopgenerics: Methods and Classes for the OMOP Common Data Model*. R package version 0.2.2, https://darwin-eu-dev.github.io/omopgenerics/.
4. Black A, Gorbachev A, Burn E, Catala Sabate M (2024). *CDMConnector: Connect to an OMOP Common Data Model*. https://darwin-eu.github.io/CDMConnector/, https://github.com/darwin-eu/CDMConnector.
5. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.
6. Du M, Catala M, Burn E, Mercade-Besora N, Chen X (2024). *omock: Creation of Mock Observational Medical Outcomes Partnership Common Data Model*. R package version 0.2.0, https://github.com/oxford-pharmacoepi/omock, https://oxford-pharmacoepi.github.io/omock/.