Software demonstration: CohortConstructor – an R package to support cohort building pipelines

Edward Burn[1], Núria Mercadé-Besora[1], Marta Alcalde-Herraiz[1], Mike Du[1], Yuchen Guo[1], Kim Lopez Guell[1], Xihang Chen[1], Markus Haug[2], Hiba Junaid[3], Daniel Dedman[4], Martí Català[1]

[1]Health Data Sciences, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, United Kingdom

[2]Institute of Computer Science (ICS), University of Tartu, Estonia

[3]Barts Life Sciences & Barts Bone and Joint Health, Barts Health.

[4]Clinical Practice Research Datalink (CPRD), Medicines and Healthcare product Regulatory Agency, United Kingdom.

Background

Cohorts are a fundamental building block for studies that use the OMOP CDM, identifying people who satisfy one or more inclusion criteria for a duration of time based on their clinical records. Currently cohorts are typically built using CIRCE (https://github.com/OHDSI/circe-be) which allows complex cohorts to be represented using JSON. This JSON is then converted to SQL for execution against a database containing data mapped to the OMOP CDM. CIRCE JSON can be created via the ATLAS graphical user interface (https://github.com/OHDSI/Atlas) or programmatically via the Capr R package (https://github.com/OHDSI/Capr). However, although a powerful tool for expressing and operationalising cohort definitions, the SQL generated can be cumbersome especially for complex cohort definitions. Moreover, when multiple cohorts are defined these are typically instantiated independently which can lead to duplication of work.

The CohortConstructor package provides an alternative approach to building cohorts in data mapped to the OMOP CDM. It promotes cohort building in a pipeline fashion, with creating base cohorts coming first which is then followed by manipulation of these cohorts to apply specific inclusion criteria. This package provides tools for common cohort-manipulation operations, such as restricting on patient demographics, calendar time, and/ or individuals' presence (or absence) in another cohort. Moreover, the package tracks the impact of applying these operations so as to allow for providing a detailed summary of cohort attrition.

Methods

CohortConstructor is an R package. Functionality included in CohortConstructor depends on a number of existing R packages, most notably building on top of the PatientProfiles R package (https://darwin-eu-dev.github.io/PatientProfiles/).

The package is tested primarily using the OHDSI omock R package (https://github.com/OHDSI/omock), which supports the creation of small test datasets in the OMOP CDM format. At the time of writing the package has 834 unit tests with 98% test coverage. These tests are run against duckdb in continuous integration, and have also been run against Postgres, SQL Server, Redshift, and Snowflake. These tests focus on small example data for which the desired result is known and so can be checked to be working as expected across different database management systems.

Although complete feature parity with CIRCE is not the goal, it is intended that the majority of existing OHDSI cohort definitions should also be possible to create using CohortConstructor. To check this we have replicated a range of cohorts from the OHDSI phenotype library (https://ohdsi.github.io/PhenotypeLibrary/). CohortConstructor also includes functionality not to our knowledge supported by CIRCE. For example, it allows cohort exit to be defined based on any user-defined variable.

To test the performance of the package we have created a benchmarking script in which we selected 10 phenotypes from the OHDSI library that cover a range of concept domains, entry and inclusion criteria, and cohort exit options. For both CIRCE and CDMConnector, we instantiated each of these cohorts separately and together as a set. Additionally, we measured the time taken to stratify a cohort by sex, and by sex and age (aged 50 or less, and more than 50 years old) using both multiple CIRCE JSON (one for each cohort) and via a cohort-pipeline with CohortConstructor. The code used for benchmarking is available in the following GitHub repository: https://github.com/oxford-pharmacoepi/BenchmarkCohortConstructor

Results

CohortConstructor has been released on CRAN https://cran.r-project.org/web/packages/CohortConstructor/index.html, with the source code available on GitHub https://github.com/OHDSI/CohortConstructor. The current version, version 0.2.1, provides numerous functions for creating and manipulating cohorts.

The package supports a cohort building pipeline where base cohorts are first created. These base cohorts are generally either based on concepts (where matching clinical events are identified) or based on patient demographics (for example, identifying the date when individuals in the database reach a certain age).

CohortConstructor offers a set of functions restriction-based and others focused on cohort creation. The first set includes requirements on demographics (age, sex, and prior and future observation), on dates (entry date or contributing time within a time-interval, and restricting subject's cohort entries to the first), and on other cohorts (e.g. require that subjects have an entry on another cohort in a defined time-period). Cohort creation functions include operations like joining or splitting overlapping periods, unifying cohorts within a cohort table, generating a matched cohort from a given cohort, and generating a cohort from the different intersections between the given cohorts.

With CohortConstructor we replicated the following cohorts from the OHDSI phenotype library: COVID-19 (ID 56), inpatient hospitalisation (23), new users of beta blockers nested in essential hypertension (1049), transverse myelitis (63), first major depression with no

occurrence of certain psychiatric disorder (1020), major non cardiac surgery (1289), asthma without COPD (27), endometriosis procedure (722), new fluoroquinolone users (1043), acquired neutropenia or unspecified leukopenia (213). For all cohorts the same individuals were identified and included in the cohorts under the CIRCE and CohortConstructor approaches.

Initial benchmarking results are shown below, running against a 100,000 person sample of the CPRD GOLD database on Postgres. By the time of the OHDSI symposium we plan to have run the benchmarking script across a range of different databases mapped to the OMOP CDM.

Cohort name	Tool	Time (minutes)
Acquired neutropenia or unspecified leukopenia	CohortConstructor	1.52
	CIRCE	1.85
Asthma without COPD	CohortConstructor	2.62
	CIRCE	8.33
COVID-19	CohortConstructor	1.38
	CIRCE	47.62
Endometriosis procedure	CohortConstructor	0.93
	CIRCE	0.82
First major depression	CohortConstructor	1.25
	CIRCE	> 2880
Inpatient hospitalisation	CohortConstructor	1.09
	CIRCE	11.66
Major non cardiac surgery	CohortConstructor	1.12
	CIRCE	19.35
New fluoroquinolone users	CohortConstructor	0.73
	CIRCE	2.55
New users of beta blockers nested in essential hypertension	CohortConstructor	2.93
	CIRCE	0.89
Transverse myelitis	CohortConstructor	0.91
	CIRCE	0.60

Table 1: Time taken to create each cohort by CIRCE and CohortContructor running on a 100,000 person sample of CPRD GOLD.

Tool	Time (minutes)
CIRCE	93.67
CohortConstructor	22.64

Table 2: Total time taken to create cohorts by CIRCE and CohortContructor, creating them together as a set with CohortConstructor on a 100,000 person sample of CPRD GOLD.

Tool	Time (minutes)
CIRCE	97.64
CohortConstructor	1.58

Table 3: Time taken to stratify a cohort by sex, and by sex and age (aged 50 or less, and more than 50 years old) by CIRCE and CohortConstuctor running on a 100,000 person sample of CPRD GOLD.

Conclusion

CohortConstructor provides a range of functions for manipulating and constructing cohorts from cohort tables in the database. Compared to existing approaches, CohortConstructor is particularly advantageous when creating complex cohorts and/ or when many cohorts are created that are related to the same domain in the OMOP CDM.