Enhancing Comparator Selection in OHDSI studies using Cohort Subset Operations: A Software Demo of the CohortGenerator R HADES Package

James P Gilbert[1], Anthony Sena[1,2], Justin Bohn[1], Christopher Knoll[1], David M. Kern[1]

[1]Janssen Research and Development, Titusville, NJ,

[2] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands

Background

Producing meaningful phenotype algorithms for sub-populations is a core goal of the OHDSI HADES R tools [1]. However, existing methods lack a consistent and programmatic approach for generating sub-populations, often relying on manual operations. Additionally, selecting appropriate comparator exposures in comparative cohort studies has been limited by the inability to consider specific sub-populations during the selection stage (i.e. the indicated population for treatment). In this work, we introduce cohort subset definitions, an extension to the *CohortGenerator* R HADES package, enabling the efficient creation of sub-populations of base cohorts. We apply this approach to the *ComparatorSelectionExplorer* [2], improving the recommendations of potential comparators in specific contexts.

Methods

Cohort subsets extend cohort definition sets, allowing subset operations based on criteria such as patient cohorts within specified date windows, specific demographics, and observation criteria. These criteria can be combined to define subset definitions, enabling easy sub-setting of initial populations. All subset definitions can be saved and reloaded in JSON format for cross-platform implementation. We apply this approach to three large-scale claims databases: Merative Medicaid, Merative Medicare, and JMDC, using RxNorm ingredients as drug exposure cohorts. Cosine similarity scores and recommendations will be calculated using the methods described in [2]. Briefly, this approach computes a similarity between feature vectors for all covariates for demographics, presentation (conditions within 30 days of exposure), medical history (conditions more than 30 days before exposure), prior medications and the recorded visit context of the exposure (if any). To select useful comparators and protect patient privacy, only cohorts with least 1000 exposed patients are included in our analysis.

Results

In this software demo, we demonstrate the implementation of meaningful subset definitions and showcase how this utility enhances existing OHDSI packages, specifically the *ComparatorSelectionExplorer*. Users can explore the results using a shiny app. The demo emphasizes the design and implementation of cohort subset definitions in R. The computational efficiency of generating subset cohorts is significantly higher compared to generating base cohorts, and the subset cohorts seamlessly integrate into existing HADES analysis tools without requiring additional modifications. Example R code for the subset definitions is shown in Figure 1. This definition can be re-used and re-applied efficiently to many cohorts.

Table 1 shows the number of potential comparators for the fluoroquinolone (broad spectrum antibiotic) levofloxacin, which is prescribed for both urinary tract infections (UTIs) and respiratory infections. In both cases the search space is greatly limited. There is between a 33% - 65% reduction and between a 38% and 52% reduction in ingredients for the UTI Pneumonia indicated populations, respectively.

Table 2 shows the ranking suggestion of comparators in context specific and non-context specific settings for levofloxacin. While the top comparator for each population is an antibiotic, the type of antibiotic

differs for the base population and each subset, highlighting the importance of considering the specific indication being studied when choosing the most appropriate comparator. It is also of note that the analysis package captures other ingredients that are likely used in combination with antibiotics, or because of the underlying populations. For example, glucose and sodium chloride are RxNorm ingredients that occur in many drugs. Similarly, confounded factors such as the high prevalence of UTIs amongst diabetics [3] may influence the resulting list when considering exposed sub-populations.

Conclusion

This software demo will showcase the new cohort subset functionality for the *CohortGenerator* R HADES package [4] by subsetting all RxNorm ingredients to contain only subjects that have previously been diagnosed with a condition indicated for by a target medication. The approach taken to creating subset definitions is in a machine-readable format that can easily be read by other programming languages and software tools. This significantly aids the reproducibility of research by replacing hand generated cohorts, which may be prone to error, or non-standardized programmatic approaches that cannot be re-used in different studies without manual changes.

However, one notable limitation of our work is that the current approach requires the use of R programming skills to define subset definitions. This likely limits the availability and accessibility of this method for users. We leave the development of user interfaces to design and save sub-setting definitions for future work.

Data Source	Base RxNorm	UTI indication subset	Pneumonia indication subset
MDCR	1027	692	630
MDCD	1172	789	656
JMDC	1114	381	530

Table 1. Per data source count of potential comparators for RxNorm ingredients with at least 1000 patient exposures.

Rank	Base population only	UTI indication subset	Pneumonia indication subset
1	ceftaroline fosamil (.867)	ceftriaxone (.91)	cefotiam (.92)
2	pantoprazole (.862)	sodium chloride (.895)	ambroxol (.917)
3	promazine (.862)	enoxaparin (.894)	bromhexine (.912)
4	ceftriaxone (.85)	cefotiam (.894)	carbocysteine (.91)
5	methylprednisolone (.85)	glucose (.894)	prednisone (.90)

Table 2. Ranking of the top 5 comparators for Levofloxacin and their cosine similarity scores (in brackets) with base populations and those with computed subset populations for Urinary Tract Infections and Pneumonia. Cosine similarity calculations are described in [2].

```r
library(CohortGenerator)

utiSubsetDefinition <- createCohortSubsetDefinition(
  name = "uti cohort subset",
  definitionId = 2,
  # The id expression is customisable - these are the resulting ids
  identifierExpression = "targetId * 100 + definitionId",
  subsetOperators = list(
    createCohortSubset(
      # The cohort id must be included in the cohort definition set
      cohortIds = 1782155,
      cohortCombinationOperator = "any",
      # This results in subjects that do not meet the criteria
      negate = FALSE,
      # These windows relate to the index of the cohort
      startWindow = createSubsetCohortWindow(startDay = -365,
                                             endDay = 0,
                                             targetAnchor = "cohortStart"),
      endWindow = createSubsetCohortWindow(startDay = -99999,
                                           endDay = 99999,
                                           targetAnchor = "cohortEnd")
    )
  )
)
```

Figure 1. Example cohort subsetting definition code to subset any cohort to include only patients that experience a Urinary Tract Infection (UTI) within 365 days of cohort index. In principle this can be seen as an indicated condition and, in this example, was applied to all RxNorm ingredients to capture context specific exposed sub populations.

References

1. Health Analytics Data-to-Evidence Suite (HADES). Available from: https://ohdsi.github.io/Hades/
2. Bohn J, Gilbert JP, Knoll C, Kern DM, Ryan PB. Large-scale empirical identification of candidate comparators for pharmacoepidemiological studies. medRxiv. 2023:2023-02.
3. Fu AZ, Iglay K, Qiu Y, Engel S, Shankar R, Brodovicz K. Risk characterization for urinary tract infections in subjects with newly diagnosed type 2 diabetes. Journal of Diabetes and its Complications. 2014 Nov 1;28(6):805-10.
4. CohortGenerator R package. Available from: https://ohdsi.github.io/CohortGenerator/