ohdsitargets – An R package for building OHDSI study pipelines using targets

Adam Black, Martin Lavallee, Asieh Golozar and Gregory Klebanov

Background

OHDSI studies are powerful means of generating real-world evidence across a network of observational databases in the OMOP format. The value of large-scale standardized analytics in improving patients' health and clinical decision-making is widely established (1). However, there are many research questions that cannot be answered with the current standardized analytic tools highlighting the need for an extendable approach to OHDSI studies. An OHDSI study involves the execution of a series of interdependent steps making it a good fit for a pipeline tool. Most OHDSI studies are written in the R programming language which has historically not had a pipeline tool. *targets* is a pipeline tool for R that aims to improve the construction of reproducible workflows by clearly defining the high-level steps using a dependency graph to illustrate which steps are up-to-date and which have been effected by changes in the analytical pipeline (2). *targets* automatically identifies dependencies between computational tasks, keeps track of which outputs are up to date, and upon changes to inputs only reruns computation that is out of date. In addition, *targets* supports the parallelization of tasks across a cluster of compute nodes. The Targetopia (3) is a collection of R packages, each supporting the use of *targets* in a specific domain by providing target factories. A target factory is a single function that groups together a collection of individual outputs allowing users of *targets* to create a large number of tracked outputs with a single function call. We introduce an R package called *ohdsitargets* which helps developers build OHDSI analytical pipelines using the *targets* pipeline tool. Through *ohdsitargets* we provide a set of target factories that are required for beginning any analysis on OMOP data in R. Further we will present a series of novel factories that fall outside the analytical tools currently available in the OHDSI community. The intent behind *ohdsitargets* is explore the benefits of using the *targets* pipeline tool for executing OHDSI studies and provide insight into development of new directions of epidemiological studies with OMOP data.

Methods

ohdsitargets is an OHDSI-specific extension of the existing *targets* ecosystem analogous to other Targetopia packages (e.g. *stantargets*) used to build make-like pipelines in R. *ohdsitargets* separates a study into a large set of granular tasks, where each task is referred to as a "target", that can be tracked for changes and parallelized across compute nodes. Since *targets* can only track static files or R objects all intermediate steps in a study must be represented as either a file or an R object. While this works well when all data is in local files or in R it needs to be adapted when creating data in a remote database, a side effect that is necessary for OHDSI studies but discouraged when using *targets* (6). To use *targets* effectively when side effects are necessary, *ohdsitargets* follows a guiding principle: Functions that create persistent data in the database must return an R object that acts as a reference or pointer to the newly created data in the database. Any function that makes use of persistent data in the database must accept as input the R object that references the data in the database that is used. If the remote data is ever refreshed, then the R object must change to trigger all downstream dependencies to refresh as well. Our aim is to minimize as much as possible the shared state during study execution and to explicitly pass data only through function inputs and outputs so that targets can track dependencies between all steps in an OHDSI study.

Results

We trial *ohdsitargets* through a "Hello OHDSI" example from the Book of OHDSI (comparing the risk of gastrointestinal bleeding between diclofenac and celecoxib (4)) using the Eunomia OMOP test dataset (5). In figure 1, we provide an example of a dependency graph dynamically generated from the *ohdsitargets* framework (cohort 1 is for diclofenac, cohort 2 is for celecoxib and cohort 3 is for gastrointestinal bleeding). The dependency graph in Figure 1 shows which targets are up to date and how they rely on each other, something that does not currently exist in OHDSI studies. Furthermore, if a single cohort file is changed then only the generation and analyses steps involving the modified cohort will be re-executed potentially saving a significant amount of computation time. This result is a proof of concept that we intend to expand upon with new analytic factories such as for time to event analysis, prevalence, and oncology characterization.

Figure 1: Dynamic Dependency Diagram using ohdsitargets
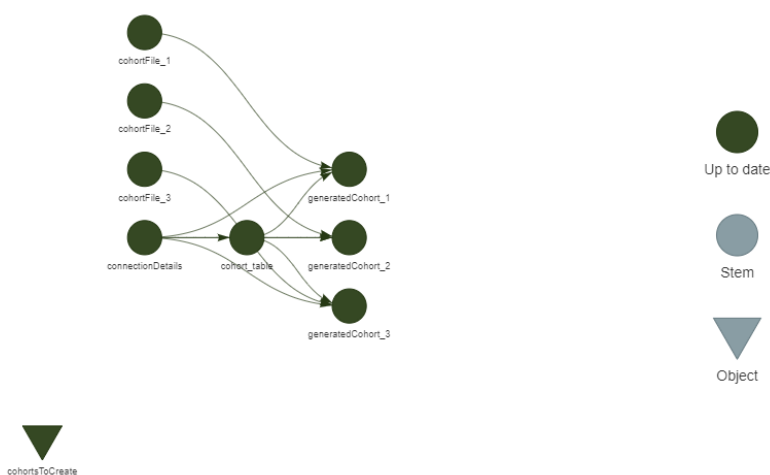
Conclusion

targets is a framework for constructing reproducible analytic pipelines in R that provides significant computational benefits. *ohdsitargets* is a contribution to the Targetopia that aims to take full advantage of the *targets* framework for OHDSI studies. We demonstrate an approach to using targets in OHDSI studies that tracks and parallelizes steps as at a granular task level and minimizes shared state and side effects in the execution of studies. In addition to standardized analyses provided in *ohdsitargets*, this framework provides the flexibility for R programmers to plug additional steps into a targets pipeline by writing an R function and adding it to the pipeline definition.

References/Citations

1. Hripcsak G, Suchard MA, Shea S, Chen R, You SC, Pratt N, et al. Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension. JAMA Intern Med. 2020 Apr 1;180(4):542–51.
2. Landau WM. The targets R package: a dynamic Make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. J Open Source Softw. 2021;6(57):2959.
3. Will Landau. The R Targetopia [Internet]. R Targetopia. 2020. Available from: https://wlandau.github.io/targetopia/packages.html
4. Martijn Schuemie, David Madigan, Marc Suchard, Patrick Ryan. Population Level Estimation. In: The

Book of OHDSI [Internet]. Available from: https://ohdsi.github.io/TheBookOfOhdsi/PopulationLevelEstimation.html#exercises-7

5. DeFalco F, Schuemie M. Eunomia: A Standard Dataset in the OMOP Common Data Model [Internet]. 2022. Available from: https://github.com/OHDSI/Eunomia

6. Will Landau. Side Effects [Internet]. 2020. Available from: https://books.ropensci.org/targets/targets.html#side-effects