

# CohortOperations: A Modular Web Tool for Enhanced Cohort Analysis on the OMOP-CDM

Javier Gracia-Tabuenca, Harri Siirtola, Anastasia Kytölä, FinnGen, Mary Pat Reeve  
Tampere University, Tampere University, Helsinki University, FinnGen, Broad Institute,

## Background

Atlas is a powerful visual tool for cohort creation, but users find it limited or non-intuitive for cohort manipulation, analysis design, and results visualization <sup>1</sup>. Hades packages create intuitive results visualizations, but cohort manipulation or analysis design requires coding skills <sup>2</sup>. In FinnGen <sup>3</sup>, we have filled this gap by developing a web tool that eases cohort manipulation, analysis design. Thanks to this, users can go all the way from cohort creation to results visualization without the need of coding skills. This first version of the tool is aimed at local analysis rather than federated analysis, but we believe it can be easily adapted to work in both modes in the future.

## Methods

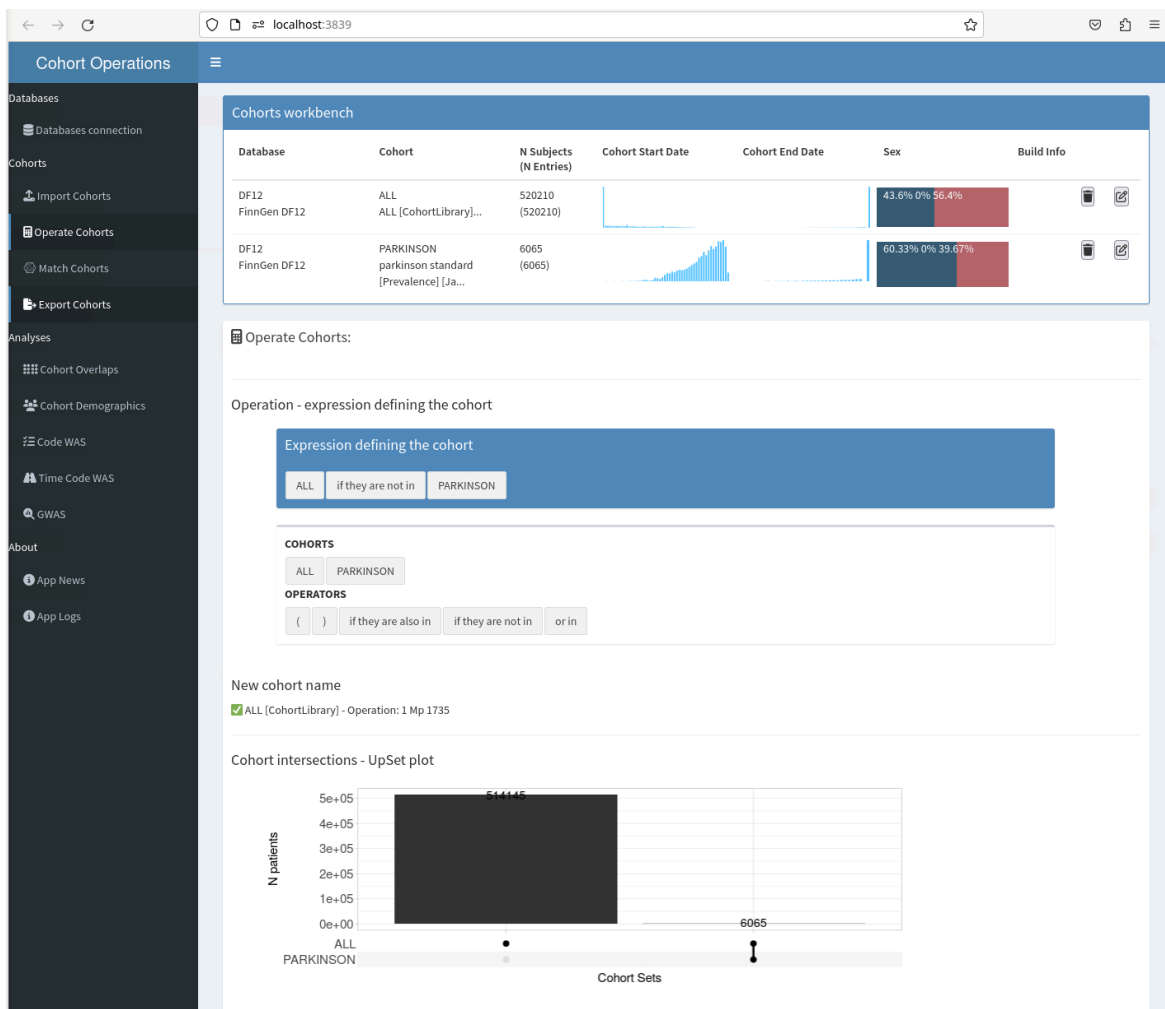
The tool is a Shiny application, named CohortOperations, that connects to one or more OMOP-CDM databases, and creates a *cohort* table. This table can be populated by importing cohorts from Atlas, a text file, or another *cohort* table. Additionally, new cohorts can be created by combining cohorts through set operations or matching. Different analyses can be selected from a list. A selected analysis will show a form asking to select one or more of the imported or created cohorts and other analysis parameters. Once parameters are selected, analysis can be run and results opened in an external app for visualization. New analysis can be included into the CohortOperations' analysis list through a yaml configuration file and external R packages <sup>5</sup>. Analysis are design in a modular fashion where each analysis module requires the following: a pair ui, server shiny functions defining the form to select the analysis parameters; an R function to execute the analysis and save the results in a SQLite or DuckDB database; and a url to an results visualization app, which can be an other shiny app or another technology. The main app and analysis modules utilize Hades package functions whenever possible, such as DatabaseConnector for database communication, CohortGenerator for cohort table management, and ShinyModules for visualizations. Functions not available in Hades have been compiled into an R package named HadesExtras <sup>6</sup>. These additional functions adhere to Hades principles and can be used alongside other Hades packages. Development and testing of these packages were conducted using Eunomia and BigQuery, with the code available on GitHub under an MIT license.

## Results

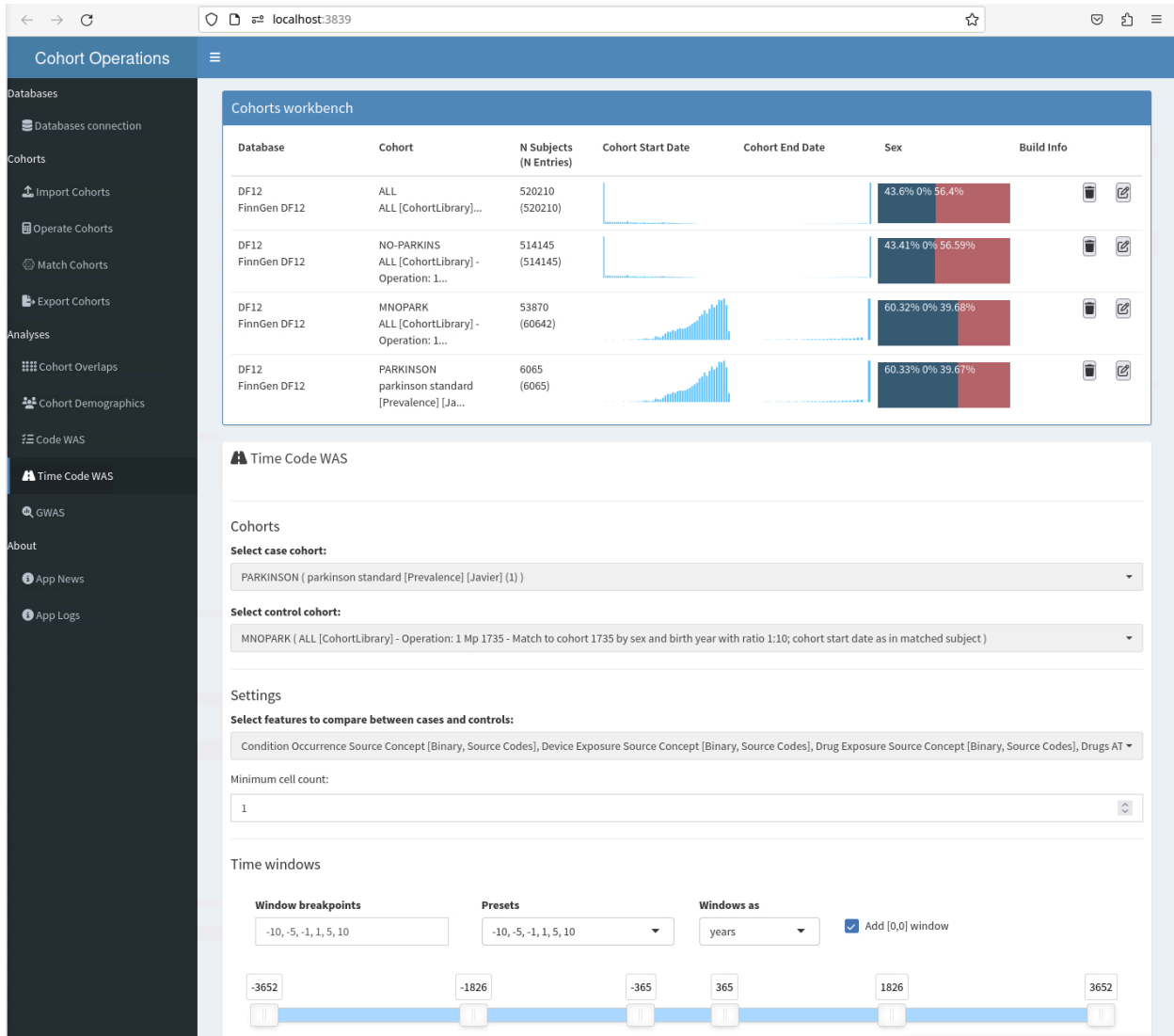
Currently, the app offers the following features: importing cohorts from an Atlas instance, a CSV file containing at least the *person\_source\_value* column, or cohorts stored in the OMOP database as *cohort* and *cohort\_definition* tables, creating new cohorts by set operations on *subject\_id*, and creating new cohorts from subjects in a target cohort with matching sex, birth year, and *cohort\_start\_date*. Additionally, created analysis modules include CohortDiagnostics, cohort overlap analysis, sex-age-onset year comparison, code-wide association studies (codeWAS), time-based codeWAS (timeCodeWAS), and genome-wide association studies (GWAS). The codeWAS calculates the statistical relevance of the prevalence of all codes in the database between case and control cohorts. TimeCodeWAS performs a similar analysis but within time windows around the *cohort\_start\_date* of case and control cohorts.

GWAS is customized for FinnGen, running on additional non-standard genetic data.

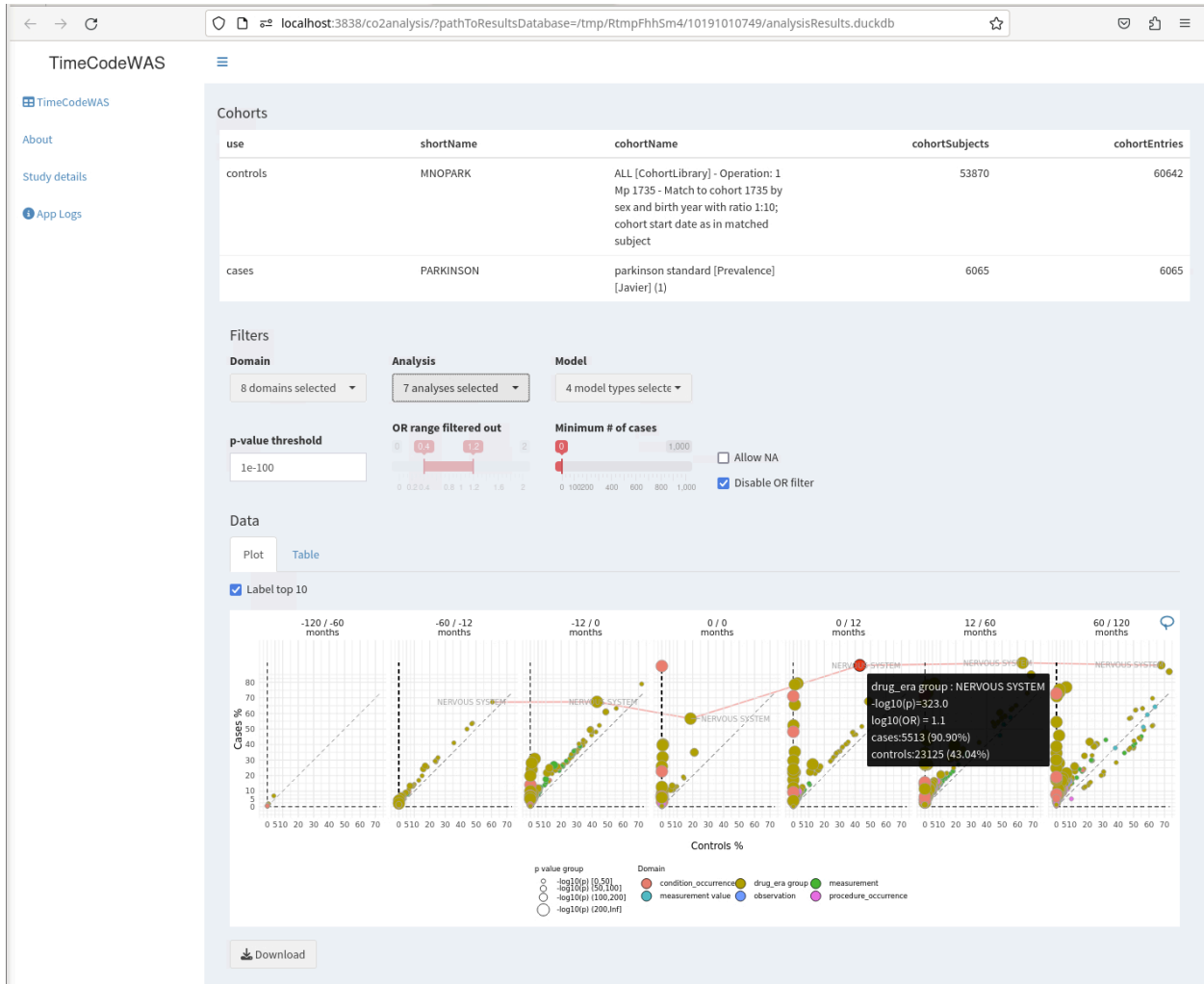
As an illustration, we executed one of the many possible workflows and captured three snapshots along the way (Figures 1 to 3). In CohortOperations, in the 'Database Connection' tab, we connect to our internal database 'FinnGen-DF12'. In the 'Import Cohorts' tab, we imported two cohorts: one from Atlas, named PARKINSON, comprising subjects with Parkinson's disease starting at onset, and the second from the precomputed library, named ALL, which includes all patients in the database starting at their first observation. In the 'Cohort Operations' tab, we created a new cohort, named NO-PARKINS, as the set difference of ALL and PARKINSON (Figure 1). In the 'Match Cohorts' tab, we created another cohort, named MNOPARK, by matching 10 subjects from NO-PARKINS for each subject in PARKINSON based on the same sex, birth year, and similar cohort start dates. In the 'TimeCodeWAS' tab, we selected PARKINSON as the case cohort and MNOPARK as the control cohort, setting windows of 1, 5, and 10 years duration before and after the cohort start date (Figure 2). After running the analysis, the results were opened in CohortOperationsViewer (Figure 3).



**Figure 1.** Screenshot of CohortOperations' 'Operate Cohorts' tab. Upper box 'Cohort Workbench', shows a table with the two cohorts that have been imported in previous steps, PARKINSON with over 6k subjects and ALL with over 520k subjects. Lower box shows a blue box where cohort names and operators can be drag and drop to build a sentence with desired operated cohort. In this case, new cohort will be "subjects in ALL if they are not in PARKINSON".



**Figure 2.** Screenshot of CohortOperations' 'TimeCodeWAS' tab. Upper box 'Cohort Workbench', shows a table with the same two cohorts that in Figure 1, plus, the cohort created after the operation in Figure 1, named NOPARKINSO, and after performing the matching, named MNOPARK. Notice how, the number of subjects in NOPARKINSO is the same as the difference of ALL and PARKINSON. Also, notice how MNOPARK has almost 10x the number of subjects and same sex and cohort start date distributions that in PARKINSON. Lower box show the options selected for TimeCodeWAS analysis.



**Figure 3.** Screenshot of CohortOperationsViewer' 'TimeCodeWAS' tab. The upper part of the visualisation shows some filters and options. The lower part show a plot with one axes for each time window, defined in month intervals. Where, each point is a medical code, x-axes are the percentage of patients with the code in controls and y-axes are the percentage of patients with the code in cases. Hovering shows details for each point.

## Conclusion

Our modular web tool bridges the gap between Atlas and Hades, enhancing cohort analysis usability for users with varying coding expertise. By enabling comprehensive and intuitive analysis and visualization, we expect this tool to significantly benefit the OHDSI community, facilitating more efficient and effective studies. Future work will focus on expanding capabilities, improving testing, and integrating additional analyses. We also invite the community to explore the tool and contribute new analyses or features.

## References

1. OHDSI/Atlas. GitHub. Available at: <https://github.com/OHDSI/Atlas>
2. Schuemie M, Reps J, Black A, Defalco F, Evans L, Fridgeirsson E, Gilbert JP, Knoll C, Lavallee M, Rao GA, Rijnbeek P, Sadowski K, Sena A, Swerdel J, Williams RD, Suchard M. *Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research*. Stud Health Technol Inform. 2024 Jan 25;310:966-970. doi: 10.3233/SHTI231108.
3. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023 Jan 18;613:508-518.
4. FinnGen/CohortOperations2. GitHub. Available at: <https://github.com/FINNGEN/CohortOperations2>
5. FinnGen/CO2AnalysisModules. GitHub. Available at: <https://github.com/FINNGEN/CO2AnalysisModules>
6. FinnGen/HadesExtras. GitHub. Available at: <https://github.com/FINNGEN/HadesExtras>