MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231108

Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research

Martijn SCHUEMIE[a,b,c,1], Jenna REPS[a,b,d], Adam BLACK[a,e], Frank DeFALCO[a,b], Lee EVANS[a,f], Egill FRIDGEIRSSON[a,d], James P. GILBERT[a,b], Chris KNOLL[a,b], Martin LAVALLEE[a,g], Gowtham A. RAO[a,b], Peter RIJNBEEK[a,d], Katy SADOWSKI[a,h], Anthony SENA[a,b,d], Joel SWERDEL[a,b], Ross D. WILLIAMS[a,d] and Marc SUCHARD[a,c,i]

[a] *Observational Health Data Science and Informatics, New York, NY, USA*
[b] *Observational Health Data Analytics, Johnson & Johnson, Titusville, NJ, USA*
[c] *Department of Biostatistics, UCLA, Los Angeles, CA, USA*
[d] *Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands*
[e] *Odysseus Data Services Inc., Cambridge, MA, USA*
[f] *LTS Computing LLC, West Chester, PA, USA*
[g] *Virginia Commonwealth University, Richmond, VA, USA*
[h] *TrialSpark Inc., New York, NY, USA*
[i] *VA Informatics and Computing Infrastructure, Department of Veterans Affairs, Salt Lake City, UT, USA*

Abstract. The Health-Analytics Data to Evidence Suite (HADES) is an open-source software collection developed by Observational Health Data Sciences and Informatics (OHDSI). It executes directly against healthcare data such as electronic health records and administrative claims, that have been converted to the Observational Medical Outcomes Partnership (OMOP) Common Data Model. Using advanced analytics, HADES performs characterization, population-level causal effect estimation, and patient-level prediction, potentially across a federated data network, allowing patient-level data to remain locally while only aggregated statistics are shared. Designed to run across a wide array of technical environments, including different operating systems and database platforms, HADES uses continuous integration with a large set of unit tests to maintain reliability. HADES implements OHDSI best practices, and is used in almost all published OHDSI studies, including some that have directly informed regulatory decisions.

Keywords. Observational research, software, open-source, machine learning, epidemiology

[1]Corresponding Author: Martijn Schuemie, email: schuemie@ohdsi.org.

1. Introduction

OHDSI (Observational Health Data Sciences and Informatics), pronounced 'Odyssey,' is a collaborative effort aiming to extract value from health data through large-scale analytics [1]. OHDSI utilizes diverse health data sources, like electronic health records and administrative claims, transformed into the OMOP Common Data Model (CDM) [2]. To analyze and generate evidence for clinical decisions, OHDSI has created HADES (Health-Analytics Data to Evidence Suite), an open-source software set used in numerous studies, some of which have influenced regulatory choices. HADES' goal is to facilitate observational research within the OHDSI community by offering a cohesive set of open-source analytic tools for characterization, causal effect estimation, and patient-level prediction. This paper outlines HADES' principles, architecture, packages, and development and adoption metrics.

2. Methods

2.1. Principles

We have developed HADES following these broad principles:

- **Open Science**: All components are open source, promoting transparency and reproducibility.
- **Direct OMOP CDM Execution**: No data preparation needed, making it versatile across diverse healthcare systems.
- **OHDSI Best Practices**: Informed by OHDSI methods research, such as supporting large-scale negative controls and empirical calibration [3,4].
- **High-Quality Software**: Documented, maintained, tested, and validated regularly.
- **Scalable Analytics**: Handles multiple questions in one analysis, even on vast datasets.
- **Big Data Support**: Operates on datasets exceeding 100 million lives.
- **Federated Analyses**: Conduct studies across OHDSI network with local patient data and shared summaries.
- **Technical Versatility**: Works on various systems and databases.

2.2. Architecture

HADES is realized through R packages, employing C++, Java, and Python for advanced analytics. For example, its core regression engine, **Cyclops**, optimizes regression models in C++, handling large-scale datasets [6]. SQL manages data manipulations, and is translated to a wide variety of platforms, while shiny [7] apps disseminate outcomes. Some HADES packages are on CRAN [8], others on GitHub.
To safeguard patient privacy in federated networks, main packages offer privacy measures, like blinding low cell counts. Data is shared in human-reviewable CSV files.
HADES' documentation employs R's standards, roxygen2 and pkgdown, encompassing reference manuals and vignettes. Continuous integration tests, spanning Windows, MacOS, and Linux, ensure cross-system compatibility.

2.3. Cohort-related packages

Cohorts are core elements of HADES analyses, capturing individuals meeting specific criteria over a time span. They signify exposures (e.g., warfarin-exposed), outcomes (e.g., bleeding cases), or special groups (e.g., pregnant women). HADES needs cohorts as inputs, with sophisticated logic managed by its packages: **Capr** for crating definitions, **PhenotypeLibrary** for storing approved cohort definitions, **CirceR** for SQL/human-readable conversion, **CohortGenerator** for CDM-compatible instantiation, and **CohortDiagnostics** with **PheValuator** [9] for assessment.

2.4. Main analytics packages

Key HADES analytics packages are:

- **DataQualityDashboard** checks conformance, completeness, and plausibility through extensive tests [10].
- **PatientLevelPrediction** conforms to OHDSI's predictive model framework [5], utilizing a broad array of predictors from CDM data. It supports diverse algorithms such as regression and gradient boosting, enabling swift external validation in OHDSI network.
- **CohortMethod** applies the comparative cohort design for causal effect estimation, utilizing large-scale propensity scores (LSPS) for confounding adjustment [11,12].
- **EvidenceSynthesis** combines results from multiple databases through meta-analysis. It includes our recent statistical approach for combining Cox models when counts are low or zero [13].
- **EmpiricalCalibration** employs negative control effect estimates to enhance causal estimates, incorporating uncertainty for scientific accuracy [3,4].

3. Results

We keep no direct measures of how often the HADES packages are used. The number of downloads in the last 14 days (measured on November 30, 2022) ranges from 2 (DeepPatientLevelPrediction package) to 1,046 (SqlRender package)

3.1. Publications

To our knowledge, HADES packages feature in 38 clinical research papers and 29 methods research papers, but there are likely more.

Notable clinical works include an in-depth study on antihypertensive drugs' effectiveness and safety [14], a COVID-19 risk calculator creation [15], and safety investigation of hydroxychloroquine, cited by the EMA for their non-recommendation [16]. HADES was also used to assess adverse effects of medications on COVID-19 [17], endorsed by EMA as best practice [18].

HADES significantly impacts methods research, evaluating causal effects [19], vaccine safety surveillance [20], and our prediction model framework [5].

4. Discussion

HADES, an R package suite, leverages the globally adopted OMOP CDM for analyzing healthcare data. It transforms CDM data into diagnostics, statistics, and visuals, shaping clinical decisions. Researchers worldwide have utilized HADES in impactful studies, with open-source code for reproducibility. HADES' liberal Apache v2.0 license fosters flexibility for collaboration, modification, and sharing. Designed for federated networks, HADES prioritizes privacy by localizing data and sharing analytics.

5. Conclusions

Developed and maintained by OHDSI, HADES evolves to enhance efficiency, broaden epidemiological designs, and offer an interactive interface for easier utilization. Access HADES at: https://ohdsi.github.io/Hades/.

Acknowledgements

Marc Suchard's HADES development is partially funded by NIH grants R01 HG006139 and R01 AI153044, along with a Department of Veterans Affairs agreement. Peter Rijnbeek, Egill Fridgeirsson, and Ross Williams are supported by the European Health Data and Evidence Network (EHDEN) project, funded by Innovative Medicines Initiative 2 JU (grant No 806968) through Horizon 2020 and EFPIA.

References

[1] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.

[2] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):54-60, doi: 10.1136/amiajnl-2011-000376.

[3] Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. Proc Natl Acad Sci U S A. 2018 Mar;115(11):2571-2577, doi: 10.1073/pnas. 1708282114.

[4] Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct p-values. Stat Med. 2014 Jan;33(2):209-18, doi: 10.1002/sim.5925.

[5] Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018 Aug;25(8):969-975, doi: 10.1093/jamia/ocy032.

[6] Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Trans Model Comput Simul. 2013 Jan;23(1):10.1145/2414416.2414791, doi: 10.1145/2414416.2414791.

[7] W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges, Shiny: Web Application Framework for R, in, 2022.

[8] R.C. Team, R: A Language and Environment for Statistical Computing, in, R Foundation for Statistical Computing, Vienna, Austria, 2022

[9] Swerdel JN, Schuemie M, Murray G, Ryan PB. PheValuator 2.0: Methodological improvements for the PheValuator approach to semi-automated phenotype algorithm evaluation. J Biomed Inform. 2022 Nov;135:104177, doi: 10.1016/j.jbi.2022.104177.

[10] Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021 Sep;28(10):2251-2257, doi: 10.1093/jamia/ocab132.

[11] Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. Int J Epidemiol. 2018 Dec;47(6):2005-2014, doi: 10.1093/ije/dyy120.

[12] Zhang L, Wang Y, Schuemie MJ, Blei DM, Hripcsak G. Adjusting for indirectly measured confounding using large-scale propensity score. J Biomed Inform. 2022 Oct;134:104204, doi: 10.1016/j.jbi.2022.104204.

[13] Schuemie MJ, Chen Y, Madigan D, Suchard MA. Combining cox regressions across a heterogeneous distributed research network facing small and zero counts. Stat Methods Med Res. 2022 Mar;31(3):438-450, doi: 10.1177/09622802211060518.

[14] Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, Reich CG, Duke J, Madigan D, Hripcsak G, Ryan PB. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019 Nov;394(10211):1816-1826, doi: 10.1016/S0140-6736(19)32317-7.

[15] Williams RD, Markus AF, Yang C, Duarte-Salles T, DuVall SL, Falconer T, Jonnagaddala J, Kim C, Rho Y, Williams AE, Machado AA, An MH, Aragón M, Areia C, Burn E, Choi YH, Drakos I, Abrahão MTF, Fernández-Bertolín S, Hripcsak G, Kaas-Hansen BS, Kandukuri PL, Kors JA, Kostka K, Liaw ST, Lynch KE, Machnicki G, Matheny ME, Morales D, Nyberg F, Park RW, Prats-Uribe A, Pratt N, Rao G, Reich CG, Rivera M, Seinen T, Shoaibi A, Spotnitz ME, Steyerberg EW, Suchard MA, You SC, Zhang L, Zhou L, Ryan PB, Prieto-Alhambra D, Reps JM, Rijnbeek PR. Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network. BMC Med Res Methodol. 2022 Jan;22(1):35, doi: 10.1186/s12874-022-01505-z.

[16] Lane JCE, Weaver J, Kostka K, Duarte-Salles T, Abrahao MTF, Alghoul H, Alser O, Alshammari TM, Biedermann P, Banda JM, Burn E, Casajust P, Conover MM, Culhane AC, Davydov A, DuVall SL, Dymshyts D, Fernandez-Bertolin S, Fišter K, Hardin J, Hester L, Hripcsak G, Kaas-Hansen BS, Kent S, Khosla S, Kolovos S, Lambert CG, van der Lei J, Lynch KE, Makadia R, Margulis AV, Matheny ME, Mehta P, Morales DR, Morgan-Stewart H, Mosseveld M, Newby D, Nyberg F, Ostropolets A, Park RW, Prats-Uribe A, Rao GA, Reich C, Reps J, Rijnbeek P, Sathappan SMK, Schuemie M, Seager S, Sena AG, Shoaibi A, Spotnitz M, Suchard MA, Torre CO, Vizcaya D, Wen H, de Wilde M, Xie J, You SC, Zhang L, Zhuk O, Ryan P, Prieto-Alhambra D; OHDSI-COVID-19 consortium. Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. Lancet Rheumatol. 2020 Nov;2(11):e698-e711, doi: 10.1016/S2665-9913(20)30276-9.

[17] Morales DR, Conover MM, You SC, Pratt N, Kostka K, Duarte-Salles T, Fernández-Bertolín S, Aragón M, DuVall SL, Lynch K, Falconer T, van Bochove K, Sung C, Matheny ME, Lambert CG, Nyberg F, Alshammari TM, Williams AE, Park RW, Weaver J, Sena AG, Schuemie MJ, Rijnbeek PR, Williams RD, Lane JCE, Prats-Uribe A, Zhang L, Areia C, Krumholz HM, Prieto-Alhambra D, Ryan PB, Hripcsak G, Suchard MA. Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis. Lancet Digit Health. 2021 Feb;3(2):e98-e114, doi: 10.1016/S2589-7500(20)30289-2.

[18] European Medicines Agency, The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) Guide on Methodological Standards in Pharmacoepidemiology (Revision 8), in, Amsterdam, the Netherlands, 2021.

[19] Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, Ryan PB, Madigan D, Hripcsak G. How Confident Are We about Observational Findings in Healthcare: A Benchmark Study. Harv Data Sci Rev. 2020;2(1):10.1162/99608f92.147cc28e, doi: 10.1162/99608f92.147cc28e.

[20] Schuemie MJ, Arshad F, Pratt N, Nyberg F, Alshammari TM, Hripcsak G, Ryan P, Prieto-Alhambra D, Lai LYH, Li X, Fortin S, Minty E, Suchard MA. Vaccine Safety Surveillance Using Routinely Collected Healthcare Data-An Empirical Evaluation of Epidemiological Designs. Front Pharmacol. 2022 Jul;13:893484, doi: 10.3389/fphar.2022.893484.