Making OHDSI Tooling accessible to Researchers and Students in a HIPAA Compliant Platform

Hannah Morgan-Cooper[1], Adam Black[2], Behzad Naderalvojoud[3], Evan Minty[4] , P Desai[5]
[1] Stanford School of Medicine and Stanford Health Care, [2] Odysseus Data Services Inc, [3] Stanford School of Medicine, [4]O'Brien Institute for Public Health, Department of Medicine, University of Calgary, Canada

Background

The OHDSI Methods Library is a powerful set of open source R packages for large scale analytics. It includes population characterization[1], population-level causal effect estimation[2], and patient-level prediction [3] that can be used on Observational Health Data in the OMOP Common Data Model. However, we have observed that the set up of these packages can be challenging and is an impediment toward adoption of OMOP. At Stanford, the de-identified Electronic Health Records (EHR) from the three Stanford hospitals and clinics is available to all researchers, pre-IRB in the OMOP Common Data Model (STARR-OMOP) via a HIPAA compliant big data computing platform. We are now making a concerted effort to make the OHDSI R packages more accessible to students and researchers who may not have a strong foundation in data or software engineering, to lower the barrier of entry to doing clinical data science with the standard OHDSI tools.

Methods

Stanford aims to provide the OHDSI tools and R Packages in its HIPAA compliant platform in a 'ready-to-run' state. As a result, we are collaborating with the extended community to solve multiple issues related to installing the OHDSI methods library R packages and their dependencies, ensuring access to the OMOP database from the hosted analytic platform called Nero, and testing OHDSI tools on Google Big Query. The end goal of this project is to run OHDSI's HADES packages and have a shareable environment in which all package dependencies are met.

Our HIPAA compliant platform, Nero, is hosted on Google Cloud Platform (GCP). This platform is approved for all high risk data and only uses products covered by the Stanford Business Associate Agreement that are contained in Google's HIPAA list. The STARR-OMOP dataset resides on BigQuery (Google's fully-managed cloud data warehouse) in its own GCP projects, and users have only read access to the STARR-OMOP data via a virtual private cloud (VPC) bridge from their own Nero GCP project. While this setup allows our data and compute resources to remain secure, the VPC firewalls often create significant obstacles for database connectivity. Nero bypasses these obstacles since all users are already in a private network. Furthermore, since Stanford is one of the few institutions with their data in the BQ database, we have had to work with Odysseus and Google to modify the OHDSI database connector to work with BQ across different GCP projects. We also provide a script to generate the connection

string for the Database Connector package. Fig 1 is a schematic representation of user access from their Nero GCP project to access the STARR-OMOP database.

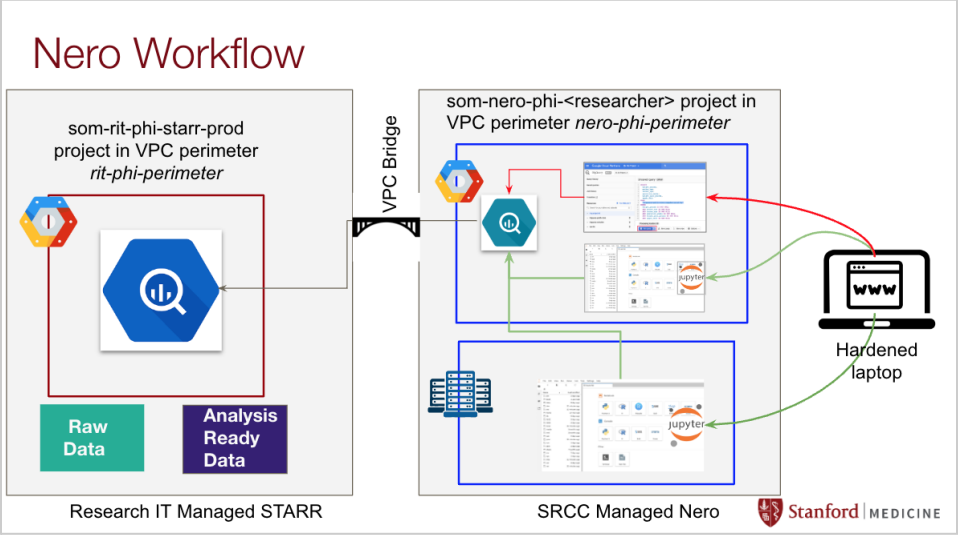

Figure 1. Schematic outlining the HIPAA compliant Nero platform hosted on Google GCP used at Stanford school of Medicine

In addition to navigating VPC firewalls and connecting to Big Query there are also a number of other set-up aspects that are challenging. In particular the installation and configuration of Java and associated technology packages, for use in R can be complicated. The Nero platform has the correct versions of technology packages installed, with all configurations set, for example, setting the path to the Java Development Kit (JDK). There will also be detailed, beginner level instructions for gcloud authentication, including how to create the json file, and set variables in the config file. Finally, all necessary variables, such as paths, will be set in the configuration file (~/.bashrc).

Furthermore, Nero provides Anaconda to manage rJava and other R packages using a Conda environment. This allows Nero users to create their own environment to run any projects developed using OHDSI methods libraries (HADES). Fig 2 describes the Conda solutions for running OHDSI studies.

The use of Anaconda for OHDSI studies not only allows users to manage R dependencies independently for each study, but it also allows users to share these environments internally. Sharing environments on Nero makes studies reproducible among users working on the same virtual machine. Because the environments are inherited from a single Anaconda, they can be accessed remotely via all Nero virtual machine instances, eliminating the need to set up a virtual machine for each study or use Docker.

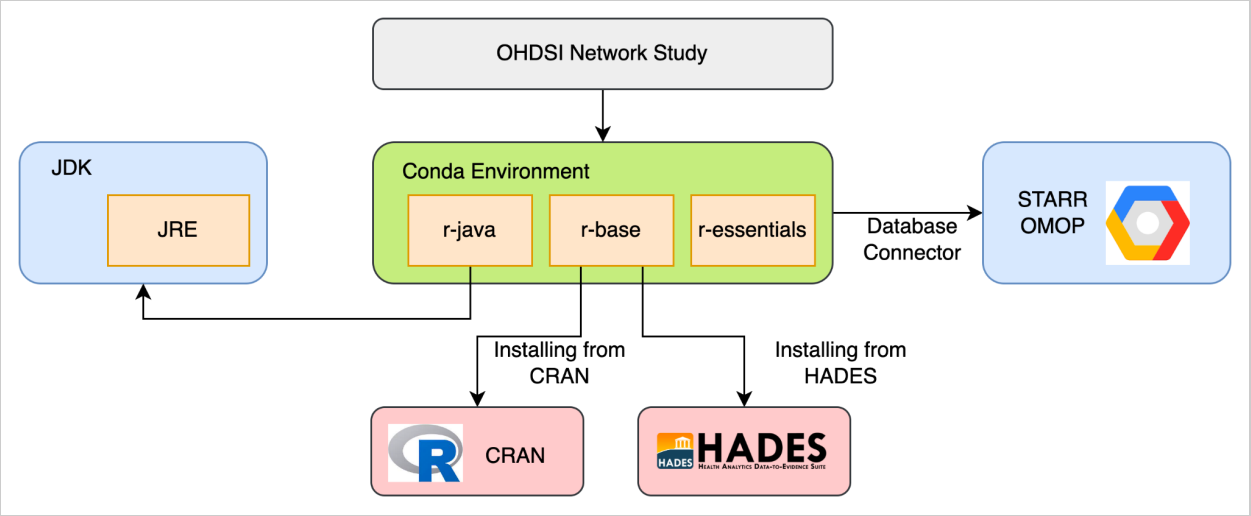

Fig 2: Schematic outlining the proposed Conda solution for running OHDSI studies

Results

Making these tools more accessible will encourage researchers to use the OMOP CDM. The platform will also allow more researchers to become familiar with the OHDSI tool set and community. Increased adoption of the CDM will lead to more collaboration between sites, since each sites' data will be compatible. Exposing more researchers to the OHDSI will also lead to more collaboration with other members of the community, and additional research studies lead by Stanford clinical scientists and researchers.

Furthermore, once implemented we will also have Conda environments for the OHDSI tool suite and Network Studies that can be shared internally.

So far the Women of OHDSI Breast Cancer and the Porpoise network studies have been run successfully using this platform.

Conclusion

Data science is a powerful tool, and has endless applications in observational health care. The OHDSI Methods library leverages data science to create tools which make performing complex analyses quicker and easier. The reproducibility and transparency of these standardized analyses makes collaboration between different institutions much simpler and easier. However, in order to fully leverage these tools the barrier for entry must be lowered since students and researchers without a background in data or software engineering may be discouraged by the complexity and difficulty of set up. The Nero platform will create an entry-way for these researchers to explore the OHDSI Methods library, and to gain familiarity with the OMOP CDM. This will further encourage research in the clinical sciences, and use of the OMOP CDM, by those who may have been discouraged by the initial barrier to entry.

References

1. Reps J, Ryan P (2023). *Characterization: Characterizations of Cohorts*. https://ohdsi.github.io/Characterization, https://github.com/OHDSI/Characterization
2. Schuemie M, Suchard M, Ryan P (2023). *CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models*. https://ohdsi.github.io/CohortMethod, https://github.com/OHDSI/CohortMethod.
3. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek P (2018). "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data." *Journal of the American Medical Informatics Association*, **25**(8), 969-975. https://doi.org/10.1093/jamia/ocy032.
4. Datta S, et al. A new paradigm for accelerating clinical data science at Stanford Medicine, arXiv:2003.10534, Mar 2020, https://arxiv.org/abs/2003.10534
5. Schuemie M, Sena A (2023). *Strategus: Coordinating and Executing Analytics Using HADES Modules*. https://ohdsi.github.io/Strategus, https://github.com/OHDSI/Strategus.