Stanford MEDICINE | Technology & Digital Solutions
STANFORD HEALTH CARE & SCHOOL OF MEDICINE

TDS Connection

THE LATEST FOR TDS PARTNERS | *May 2023*

FEATURED STORY

Power of the Commons

How Stanford Harnesses Diverse Data to Enhance Population Health Research

*By **Deepa Balraj**, Senior Software Engineer, Research Technology, TDS, and **Somalee Datta**, PhD, Director, Research Technology, TDS*

Research Technology and the Stanford Center for Population Health Sciences (PHS) are working together to improve the health and well-being of patients at Stanford using a bigger and more diverse set of information from people all over the United States. This is a unique opportunity that has never been available before.

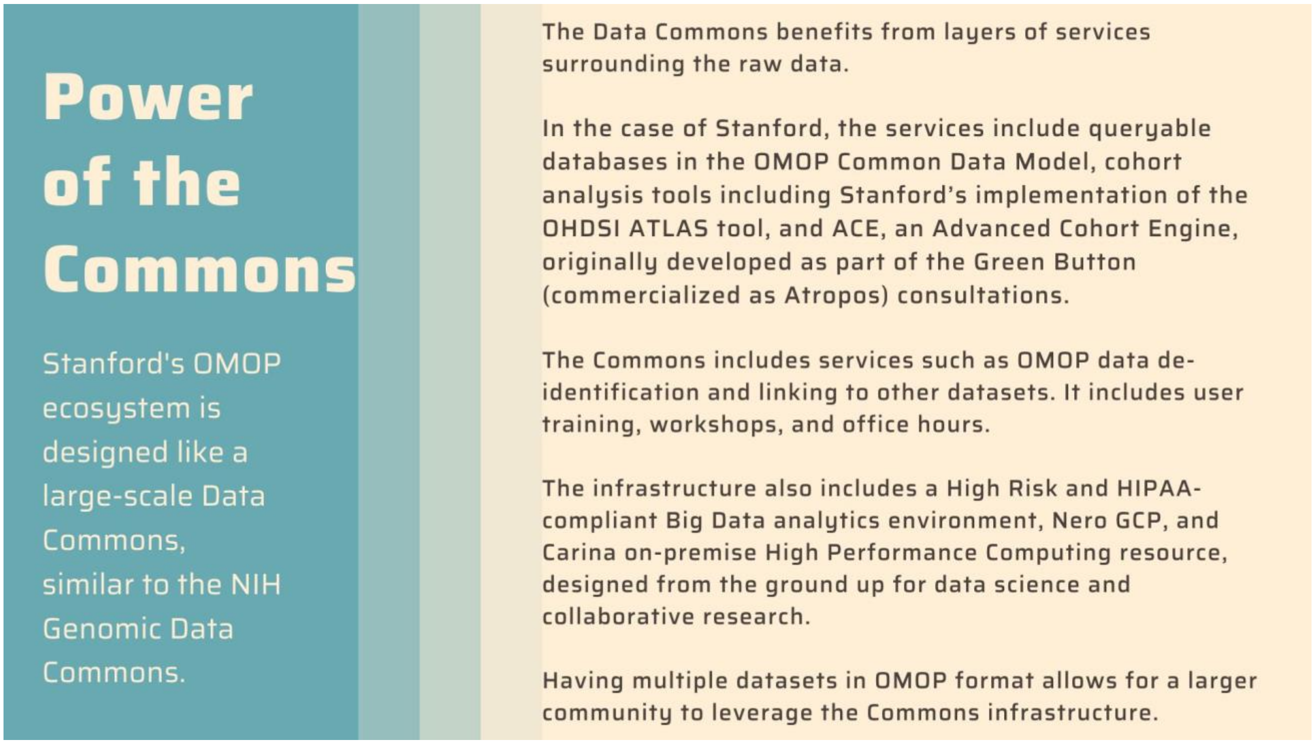
This dataset, called the American Family Cohort, represents a geographically, ethnically, and socio-economically diverse population across the U.S.

Stanford Medicine has a significant interest and advantage in developing learning healthcare systems, and the data used to train and validate these systems must represent more diversity to avoid bias that could inadvertently harm a subset of patients.

For example, if we are going to build an algorithm that detects cancer, it needs to incorporate data from patients who lack screening, as well as from patients who have access to regular screening, to be inclusive.

PHS is one of 18 cross-disciplinary centers and institutes at Stanford University, reporting to the Provost. In 2016, the Center was bequeathed an initial endowment of four new tenured faculty and a total of $18 million for its first eight years of operation.

PHS has previously provided Stanford researchers access to many high-value data presentations including Merative (formerly IBM) MarketScan, the Healthcare Cost and Utilization Project (HCUP), and Medicare. The addition of the American Family Cohort dataset is the latest jewel in Stanford Medicine's Data Science and population health toolbox.



Power of the Commons

Stanford's OMOP ecosystem is designed like a large-scale Data Commons, similar to the NIH Genomic Data Commons.

The Data Commons benefits from layers of services surrounding the raw data.

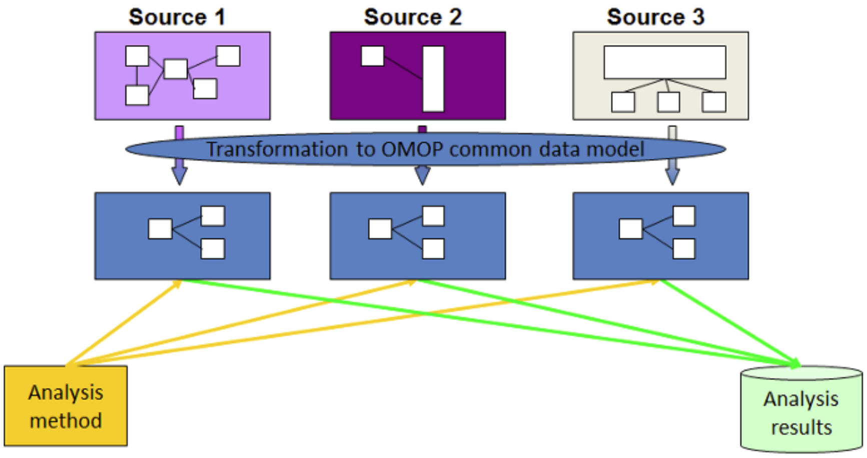
In the case of Stanford, the services include queryable databases in the OMOP Common Data Model, cohort analysis tools including Stanford's implementation of the OHDSI ATLAS tool, and ACE, an Advanced Cohort Engine, originally developed as part of the Green Button (commercialized as Atropos) consultations.

The Commons includes services such as OMOP data de-identification and linking to other datasets. It includes user training, workshops, and office hours.

The infrastructure also includes a High Risk and HIPAA-compliant Big Data analytics environment, Nero GCP, and Carina on-premise High Performance Computing resource, designed from the ground up for data science and collaborative research.

Having multiple datasets in OMOP format allows for a larger community to leverage the Commons infrastructure.

Stanford researchers now have seamless access to three large-scale observational health datasets: AFC with data counting over 7 million patient EHR, STARR with over 3 million patients' EHR data, and MarketScan with 107 million patient claims data. The three are available in a single common data model (CDM) called the Observational Medical Outcomes Partnership (OMOP) CDM, for joined analytics. The OMOP CDM) is an open community data standard that is broadly adopted by industry, government, and academia across the U.S., Europe and Asia. As of 2022, it includes over 3,000 collaborators from 400 organizations in 80 countries.



A simplified view of how a common data model works.

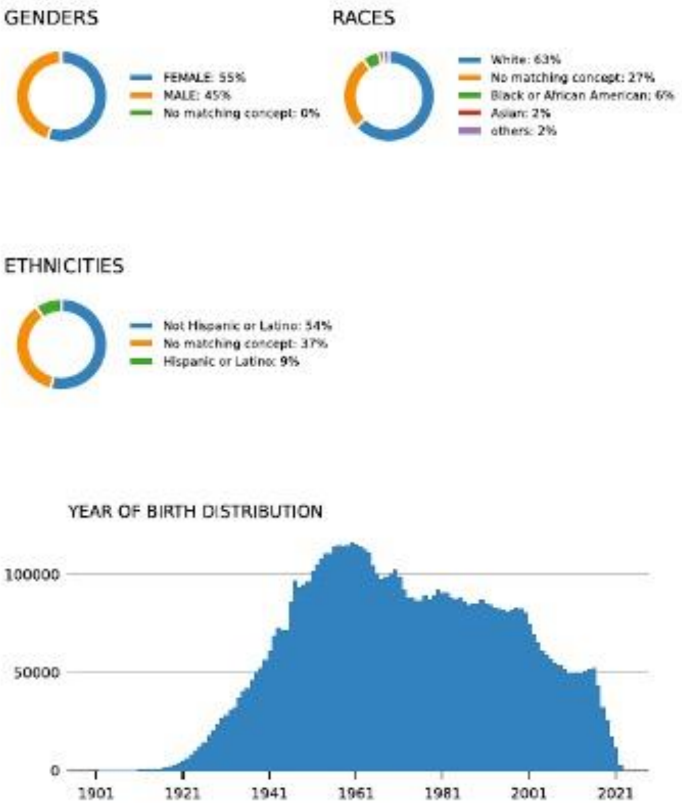
Disparate datasets, when standardized to common data models such as OMOP, significantly reduce the individual researcher's burden of data pre-processing and cleaning, thereby shortening the time to insight. Furthermore, once data is converted to OMOP format, evidence can be generated using standardized analytics tools, thereby enhancing research reproducibility.

The AFC research data are derived from the American Board of Family Medicine (ABFM) PRIME Registry. The PRIME Registry represents over 2,500 active clinicians from 50 states with data on nearly eight million patients, dating back to 2010.

The PRIME Registry is sponsored by the American Board of Family Medicine (ABFM), whose objective was to establish a Qualified Clinical Data Repository for primary care. Data were electronically extracted directly from the electronic health records (EHRs) via online portals; data elements include both structured and unstructured data, typical of disparate EHRs.

Data elements include patient demographics, diagnoses and interventions for the patients such as medications and therapies, encounter-specific data, patient-reported outcomes, area-based deprivation, and some limited clinician-specific details. Since 2019, the ABFM and the Stanford University Center for Population Health Sciences have partnered to create the American Family Cohort from the PRIME Registry data.

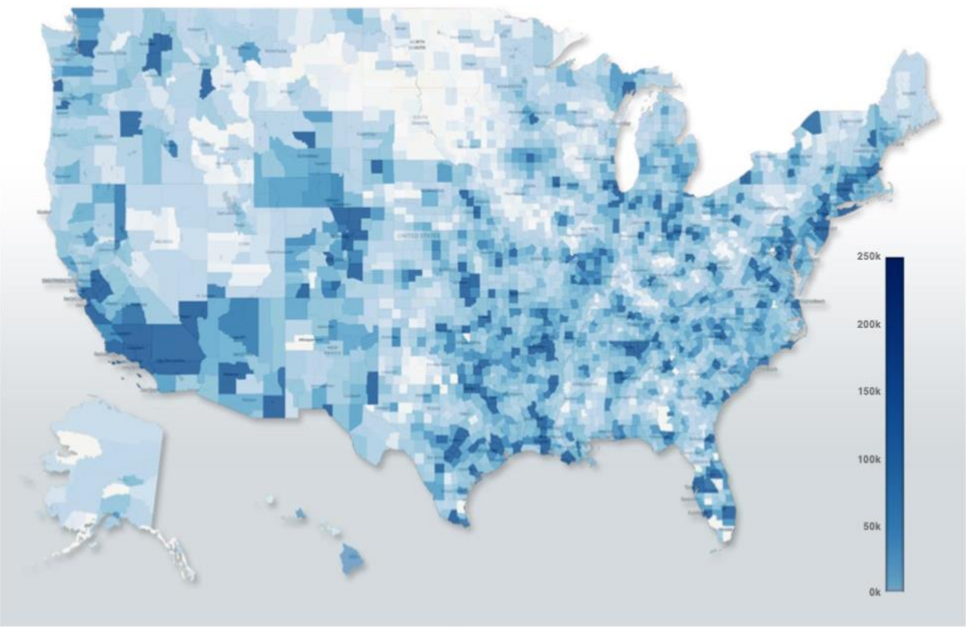
In 2022, PHS and Research Technologies entered a collaboration to convert the AFC dataset to OMOP format. We built on the success of STARR-OMOP to create AFC-OMOP, leveraging common ETL (extract, transform, and load) and



AFC-OMOP presents 7.1 million patients, 466 million conditions, 115 million procedures, 400 million observations, and 1.7 billion measurements. Shown is the distribution of gender, race, ethnicity, and year of birth.

QA (quality assurance) methodologies and cloud infrastructure to fast-track development to support an ongoing study with the U.S. Census. AFC-OMOP data access is controlled via a Google workgroup that is created and maintained in PHS data core. The data core is powered by the Redivis platform that underlies the Stanford University Library (SUL) Data Farm. The OMOP data — STARR, AFC, and MarketScan — are stored as a BigQuery datasets in a secure Stanford owned Google Cloud data center.

Figure 1: Number of AFC patients by county



The data is racially and ethnically diverse and includes 540,000 Black patients, 150,000 Asian patients, 51,000 Native American and Alaska Native patients and 16,000 Native Hawaiian and Pacific Islander patients. 4.8 million patients are white, and 758,000 patients have identified as Hispanic or Latino.

Access to individual OMOP datasets are controlled by PHS or Research Technology managed Google workgroups.

Researchers access these datasets via Stanford's HIPAA-compliant Big Data research computing platforms, the on-premise Carina and cloud-based Nero. These research computing platforms, on a Jupyter hub, are developed in collaboration with Stanford University IT Research Computing Center, PHS, and Research Technology. Notebooks with kernels for different programming languages (python, R, SAS, Stata) are available and allow for reproducible and shareable research. Tutorial videos on how to access and analyze OMOP data using Jupyter Notebooks are available on the STARR YouTube channel.

Special thanks and recognition

We acknowledge **Isabella Chu**, Associate Director of the Data Core at PHS, and **Prof. David Rehkopf**, Director of PHS, for their long-term partnership. We thank members of TDS PMO, Research Technology and Odysseus Data Services, and the OHDSI consortium who provided OMOP R&D and product development support.