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

TDS Connection

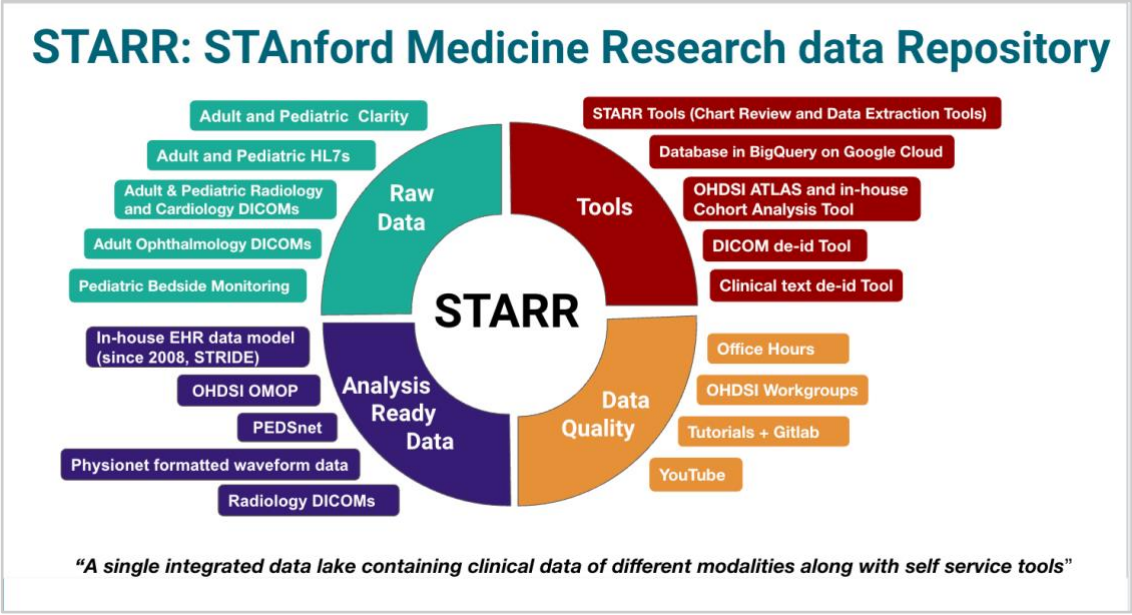
THE LATEST FOR TDS PARTNERS | *August 2022*

FEATURED STORY

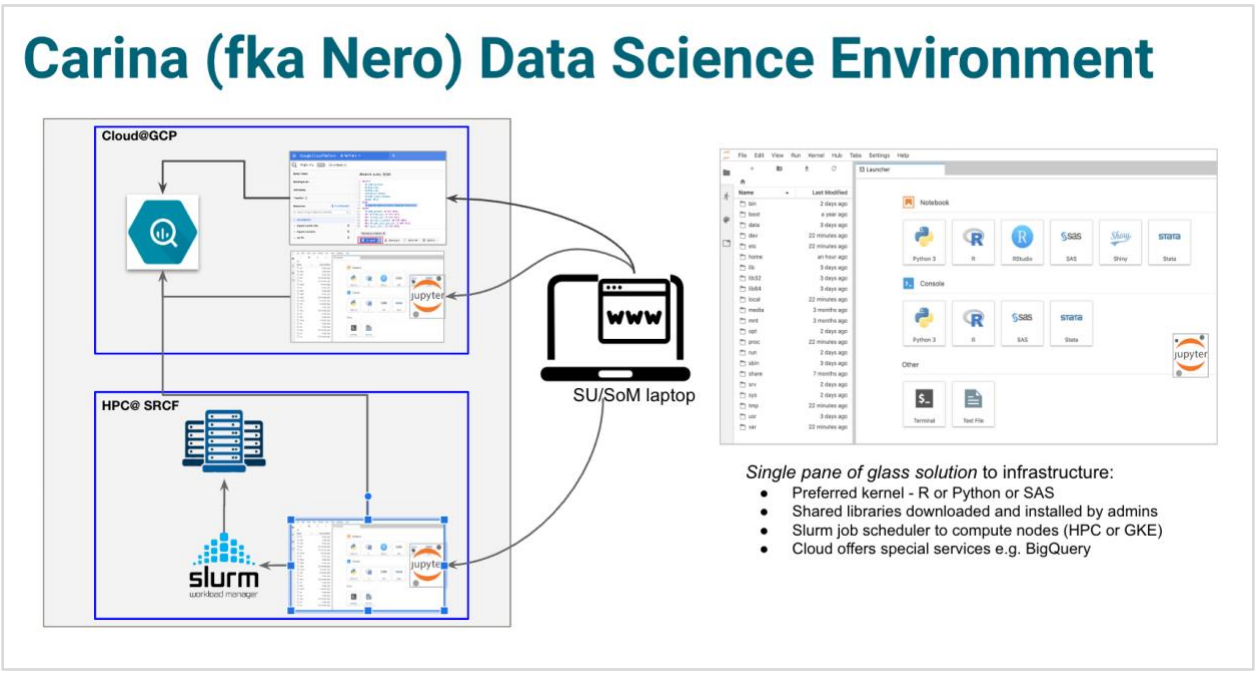
STARR and Our Journey Towards Better Data Quality and Accessibility

By Priya Desai, MS, Biomedical Informatics R&D Manager, Research IT

Clinical Informatics is at the confluence of rapid advancements in cloud computing, new techniques in data mining and machine learning, and more than a decade worth of electronic medical records data. Stanford has responded to this opportunity by creating STAnford medicine Research data Repository (or STARR, https://starr.stanford.edu/), a single integrated data lake containing clinical data of different modalities from the two hospitals – Stanford Healthcare and Stanford Children's hospital (aka LPCH). As stewards of Stanford's healthcare data, the Research IT Team builds and maintains the infrastructure for STARR, including consistently providing researchers with raw and analysis-ready data. In addition, we also provide research support, training and custom engineering services.



The STARR ecosystem is accessible via Stanford's first HIPAA compliant, Big Data computing platform, Carina (f.k.a Nero). It is built and maintained by Stanford Research Computing Center (SRCC) and has an on-premise and cloud footprint. The computing environment is a Jupyter hub which is designed to look and feel the same, whether using on-premise or cloud facilities. 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 STARR data using Jupyter Notebooks and much more are available on the STARR YouTube channel.

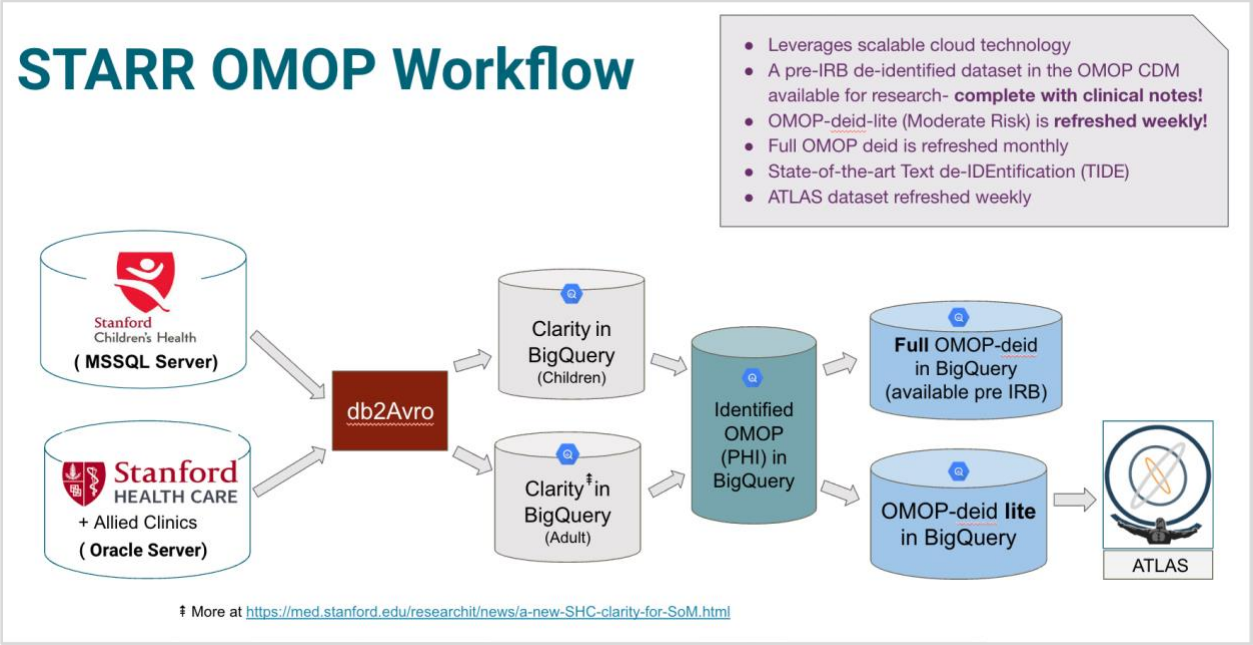


Anticipating a rapid increase in the amount of healthcare data available for research, the Research IT Team decided to leverage the scalability of cloud technologies as we established Stanford's next generation secure data center. We chose Google's cloud platform. Today, the raw EHR data (Clarity tables from both hospitals), the final downstream products like the in-house data model (f.k.a STRIDE), COHORT and Chart Review tools, the newer STARR-OMOP dataset, as well as the ETL's that convert the raw data to its analysis ready form ***are completely hosted*** on the Google cloud platform. This was largely facilitated by db-to-avro, a novel in-house solution developed to migrate large databases to the Cloud efficiently. As a result, a full copy of Clarity from both hospitals is now available on BigQuery and is refreshed daily at a regular cadence. BigQuery is Google's managed scalable database with near real-time performance on complex queries that otherwise took hours to execute on relational databases. Now we have consistent, stable, reliable data with high fidelity leading to improved data quality in the downstream products.

Utilizing the vast scalable storage capacity of the cloud, Research IT has also been able to integrate all historical radiology imaging data as well as all the Bedside Monitoring data from LPCH with STARR. That combined with the cloud pipelines developed by Research IT to reliably and efficiently de-identify DICOM imaging data, waveform numeric data, and metadata, in a fast and cost-effective manner supports Stanford researchers as they continue to push the boundaries of medical research.

In 2019, we converted the EHR data from both our hospitals and affiliated clinics into the OMOP Common Data Model. OMOP is an acronym for Observational Medical Outcomes Partnership and is maintained by the open source OHDSI consortium, an international group of scientists doing Observational research. OMOP allows users to generate evidence from a wide variety of sources in a standard data format that fosters multi-institutional collaborative research. All Stanford researchers can get access to the deidentified OMOP data pre-IRB via Carina. Our OMOP dataset includes structured EHR data as well as de identified clinical notes. As a result, many research studies focusing on algorithm development or population health studies can use these data assets ***directly without ever needing to go through IRB***, greatly enhancing accessibility to research data.

Leveraging the compute capability of the cloud, Research IT could bring sophisticated text mining and de-identification algorithms to clinical text and imaging. For de-identification, we currently use CoreNLP from Christopher Manning's lab (a Stanford CS faculty) for name entity recognition and "Hiding in Plain Sight" from Bradley Malin's lab, Vanderbilt Privacy faculty, to minimize re-identification in clinical text. We also mine the dark data hidden in clinical text and publish searchable text annotations in the NOTES_NLP table to improve cohort findings. The clinical text processing pipeline extracts term mentions i.e., string unique identifiers (SUI) within the UMLS vocabulary. The image de-identification pipeline can support de-identification of millions of DICOMs in matter of minutes. At the heart of the imaging pipeline is the popular open source MIRC-CTP framework that has been augmented to support large number of modalities and resolutions. We scrub DICOM metadata and pixel PHI.



In our commitment to improving health equity research, we've expanded the deid OMOP to include zipcode data as well as Payer information. We have zip5 data for 78.5% of all patients and payer plan coverage data including financial class and period of coverage for 95% of our SHC patient encounters. This is really exciting, as it will allow us to link to publicly available Social determinants of Health (SDoH) Datasets as well as examine correlation between SES, insurance coverage and health care utilization.

In the spirit of increasing data utilization, we have brought in the flowsheets data into the observation table. Flowsheets are a source of a lot of clinical information. Currently we have brought in the flowsheets into the observation table as json strings, and are now working on mapping concepts from it. We have already mapped the vitals found in flowsheets such as blood pressure, oxygen level, heart rate, respiratory rate, measurements from Sequential Organ Failure Assessment (SOFA) score, Glasgow Coma Scale Score, and Deterioration Index Score, to their appropriate standard concepts in the measurements table.

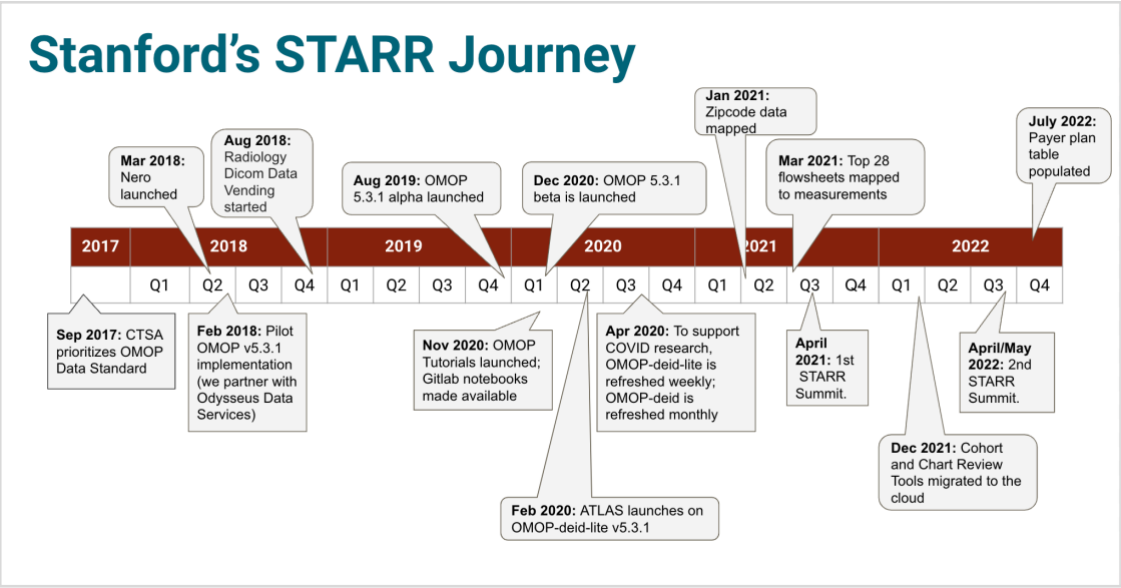
In addition to the in-house Cohort and Chart Review tool which are now hosted on the cloud, we also host the Stanford ATLAS, a free web-based tool developed by the OHDSI community to explore the structured part of the OMOP CDM. Our installation of ATLAS points to the deidentified STARR OMOP and is refreshed weekly. ATLAS allows you to create cohorts, define phenotypes, run characterization studies as well as network studies. It is a cohort tool and much more. At Stanford we enabled a lesser-known feature of ATLAS called the execution engine which greatly eases the challenges of running the network studies end to end. At this time, we have over 200 researchers who access ATLAS; and Stanford has participated in over 25 OHDSI network studies.

The onset of COVID created an increased urgency for reliable data that was updated more frequently. We needed to make rapid modifications to ETL's pipeline and this was a catalyst for us to streamline our software processes and make them more robust. We put in place data quality checks right from the first very step of getting the Clarity data into Bigquery. Checks included basic row count checks, generating dashboards, and automated email notifications to alert us if row counts dropped and rose above a certain threshold as tables were migrated or generated.

We also generated a "regression" dataset consisting of about 150 thousand MRNs from both hospitals, containing patient data that had a similar distribution to the actual patient population in terms of demographics as well as diagnosis, procedures and patients whose data we were not allowed to use. Before any ETL modifications were brought into the production pipeline, we ran the entire pipeline on the regression dataset and this has often alerted us to issues before the code went into production.

As part of improving our data quality, we stabilized person_ids, provider_ids, caresite_ids and visit_occurrence_ids in our deid data between runs. What that means is "Persons identified in their cohort stay in their cohort". This allows for ***reusability and predictability.***

The journey continues as we work towards better data quality to accelerate research.



Appendix: STARR Resources & Some Notable Achievements

STARR is a data repository that underlies a number of tools and pipelines. Your one-stop-shop for finding all STARR resources is the primary STARR website. Of particular interest are the web pages, self-service tools and consultation services.

STARR ecosystem is heavily subsidized by SoM Dean's office and the CTSA grant. However, it is not possible to subsidize 100% of everything, so we have a cost-sharing model with our research community that is updated from time to time.

Since the pandemic, OHDSI consortium's primary focus has been COVID-19 network studies. Our OMOP release timing was fortunate, we were in beta at the end of 2019. Since early 2020, Stanford has participated in over two dozen network studies. Eleven of these are now published, ten of the eleven are related to COVID-19.

We have hosted an annual STARR-Informatics Summit for two years in a row showcasing research enabled by STARR.

We are also delighted that Atropos (fka Green Button) consultation service, which was originally developed on STRIDE data extracts converted to OMOP prior to commercialization as Atropos Health. Atropos is now running an operational pilot at Stanford Health Care and is leveraging a OMOP dataset that Research IT refreshes monthly to meet Atropos prognostogram service requirements.

We recently had the opportunity to present to the CTSA community, a national network of more than 50 medical research institutions. Stanford School of Medicine is a hub in this network (https://med.stanford.edu/spectrum.html). The talk storyboard followed a recently featured blog run by CTSA coordinating center, Center for Leading Innovation and Collaboration (CLIC).