MOUNT SINAI DATA WAREHOUSE ANNUAL REPORT SEPTEMBER 23, 2021

Scientific Computing and Data

ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI

Table of Contents

Introduction

We are pleased to present this annual update on the Mount Sinai Data Warehouse (MSDW) covering the period September 1, 2020 through August 31, 2021.

A major focus of this year was the implementation of MSDW2, a re-platforming of MSDW from an Oracle database on stand-alone servers to a Microsoft SQL Server database hosted on the Minerva HPC cluster. Additionally, MSDW2 features OHDSI's Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), the new *de facto* standard for multi-institutional research data sharing.

We have also released two new patient cohort query tools, Leaf and ATLAS, for querying identified and de-identified data from MSDW2's OMOP CDM.

We continue to provide data to the Sinai user community, CDRN, and Sema4 while improving our processes and procedures.

At a high level, we follow these processes and procedures:

(1) We only share data sets with requesters listed on the IRB (rather than try to collect training information and/or training verification from PIs).
(2) We require IRB protocol and approval forms as part of our automated ticket intake process in Jira.
(3) We review the dates on IRB, quality improvement, data mart, reports and operational data request every three months and request updated documents or revoke access to data.
(4) We continue to update and provide a de-identified COVID-19 data set to make COVID-19 data maximally available.
(5) We follow a "Caboodle-first" policy, where we fulfill requests by extracting data from Caboodle where possible, then from Clarity when necessary, and from other databases as a fallback. We developed this order of precedence to maximize data quality, timeliness, traceability, and labor efficiency.

Accomplishments over last year

COVID-19 De-Identified Data Sets

Starting in March 2020, the MSDW team created a de-identified COVID data set containing demographic and clinical information. The MSDW team refreshed this data set daily until switching to a weekly refresh cadence on September 6, 2021. The updated data files are stored in a secure folder for end users in the Mount Sinai community to download via the MSDW website and via Minerva Data Ark. Details of how we removed the 18 PHI elements and shifted the dates are in the IRB.

The data set comprises multiple data files. The main data file is the Patient Encounter file, which contains demographic and encounter level information for each COVID related encounter at a facility in the Mount Sinai Health System using the Epic electronic health record. The additional files (daily vital signs, medication administration, lab results, diagnosis, and radiology impressions) can all be linked to the patient encounter, as well as one another, using the masked MRN and masked encounter ID. Through the end of August 2021, the de-identified COVID-19 data sets have been downloaded more than 6,600 times by 300 unique users.

In addition to the de-identified COVID-19 data set, the MSDW team has provided 130 custom data sets for COVID-19 data.

MSDW2

The major focus for this year was the release of MSDW2, a re-platforming of MSDW to a Microsoft SQL Server database fed primarily from Caboodle and featuring the OMOP Common Data Model. This first version will cover the core set of data: patient demographics, patient social and family history, past medical and surgical history, providers, care site locations, encounters of all types, diagnoses (from both problem lists and encounters), procedures, laboratory test results, vital signs, medication orders and administrations. This core data set will be refreshed daily and will be available in both identified and de-identified versions.

MSDW2 exhibits the following features:

- Uses Mount Sinai's Epic EHR as its primary source of data (rather than HL7 interface messages).
- Situated on the Minerva High-Performance Computing (HPC) cluster, rather than on its own hardware, to lower operating costs and to facilitate links to other research data sets already housed on Minerva.
- Replaces the legacy MSDW's bespoke data model with the OHDSI consortium's standard OMOP Common Data Model to minimize the cost of building and maintaining ETL data pipelines, and to facilitate data exchange with research partners.
- Leverages, as much as possible, the features and functionality of Epic's Cogito databases (Clarity and Caboodle), as developed and supported by Epic Systems Corporation.
- Replaces the legacy MSDW's Oracle database with a Microsoft SQL Server technology stack, designed to align with Epic Caboodle and to lower software licensing costs.

Cohort Query Tools

In September 2021, CQT query tool was replaced with 2 new query tools: Leaf and ATLAS.

Leaf is a lightweight drag-and-drop query tool that enables researchers to easily create patient cohorts from de-identified data. Leaf contains de-identified data only for MSDW2. Leaf is open-source software, the development of which is led by Nicholas Dobbins at the University of Washington. Arthur Goldberg, a software engineer on the MSDW team, is now collaborating with Nic on Leaf's development.

ATLAS is web-based tool for querying OMOP databases. Users can define their own patient cohorts, browse standardized healthcare vocabularies, and conduct cohort-level analysis. ATLAS can be used to query MSDW2's de-identified data, PHI data, or data marts with approved IRB protocols. ATLAS is open-source software being developed by the OHDSI community.

Migration of MSDW Dependencies

With the launch of MSDW2 on Minerva, the MSDW team has worked to migrate all dependencies to either this new system or to Epic such that the legacy Oracle-based MSDW can be retired.

- **Reports** – All active MSDW end-user reports were migrated to source from Clarity or Caboodle. The Press Ganey reports, based on hospital billing data for the HCAHPS patient satisfaction surveys, were transitioned to the Enterprise Data Warehouse team. The Adolescent Health Center (AHC) report was transitioned to the Adolescent Health Department**.**

- CDRN quarterly data loads and COVID-19 data loads were also migrated to Clarity or Caboodle.
- **Data Marts** – In Q4 2021, the data marts for Anal Cancer, Liver, iSITE, and BioMe will be re-implemented in MSDW2 using the OMOP Common Data Model.

Data Sources

In the last year, our primary data source for custom queries and data marts has been Epic Clarity and Caboodle databases. MSDW2 will also be sourced primarily from Epic Caboodle. Other data sources include Provation MD for gastroenterology procedure and MSX data warehouse for billing and insurance data.

Increased staffing

Over the past year, we have grown our team by filling 8 open positions. These positions include a data warehouse developer, a REDCap administrator, a sematic front-end developer, a semantic back-end developer, a data discovery developer, and a technical writer. We have also added two key senior positions:

Director, HHEAR & Semantic Technology – James "Chip" Masters, PhD

Chip is an expert in semantic technologies and has a PhD in Applied Mathematics. His role is to develop the HHEAR data portal and integration with HADatAC (Human Aware Data Acquistion).

Clinical Data Strategist – Fabricio Kury, MD

Fabricio is well-qualified clinical informatics professional and physician. He has extensive experience working with the OMOP Common Data Model, the ATLAS cohort query tool, and other tools developed by the OHDSI community.

Metrics

The following table contains the Key Performance Indicators (KPI) for the MSDW team during the period covered by this report.

Metric	I2B2	TriNetX	MSDW	CQT	OMOP
# of user tickets created	30	6	292	52	10
# of user tickets closed	30	6	234	52	4
# of unique users	122	-	-	774	92
# of active unique users	13	-	-	52	-
# queries run by users	205	-	138	721	8
# active projects /data marts (last 6 months)	-	-	4	4	-
# total projects/data marts	-	-	-	24	

Data Quality

The MSDW team has put in place several measures to improve the quality of the data that is being provided to requestors. We follow a "Caboodle-first" policy, where we fulfill requests by extracting data from

Caboodle where possible, then from Clarity when necessary, and from other databases as a fallback. We developed this order of precedence to maximize data quality, timeliness, traceability, and labor efficiency.

Epic Caboodle constitutes the primary data source for the new MSDW2 system. MSDW2 further enhances the Caboodle data with additional data processing to comply with the requirements of the OMOP, including (1) reformatting the data into the structure of the OMOP Common Data Model, and (2) standardizing the content of the OMOP database by mapping all source values and categories to standard healthcare codes and terminologies (e.g., SNOMED-CT, LOINC, RxNorm, CPT4, HCPCS, ICD-10-CM). The MSDW team has designed MSDW2 to provide transparent data lineage back to its Caboodle source tables and, in some cases, further upstream to the Epic Clarity and Chronicles databases.

Collaborations

CDRN

Mount Sinai participates in INSIGHT Clinical Research Network (previously NYC-CDRN), a PCORI-funded network of seven major NYC hospital systems. INSIGHT CRN collects patient data from all sites into a common OMOP data model in order to facilitate regional and national patient-centered research. In accordance with BRANY IRB approval, MSDW provides quarterly data submissions to INSIGHT and, since March 2020, bi-weekly submissions of data relevant to COVID-19 research.

Sema4

We regularly provide de-identified or identified data sets to Sema4 through specific research projects as well as through legal agreements such as the DSCA. We provide monthly identified data sets for the JCAP agreement. Starting in October 2021, the MSDW team will provide the JCAP data set in MSDW2's new OMOP format.

We worked with the Chief Medical Informatics Officer, Bruce Darrow, to help develop a process to remove data from patients who wanted to "opt-out" of sharing data with Sema4.

Goal for the upcoming year

Our roadmap for 2022 includes the following initiatives:

- Addition of new data elements and data sources to MSDW2 including clinical notes, immunizations, patient allergies, genomics data, and billing data.
- Ingestion of DICOM metadata elements from the Imaging Research Warehouse (IRW)/XNAT for Leaf/ATLAS searches.

Conclusion

The MSDW team is grateful for the guidance from the IRB, Compliance, Legal, MSIP, IT, MSHS researchers and clinicians, and senior leadership to help us continue to improve and provide a valuable and safe service for MSHS.