MOUNT SINAI DATA WAREHOUSE ANNUAL REPORT AUGUST 31, 2022

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 from September 1, 2021 through August 31, 2022.

Our major focus this year was on prioritized enhancements to MSDW2 and our cohort query tools (Leaf and ATLAS). 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 Mount Sinai user community and to external partners, such as Insight CRN (formerly 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 requestors 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 requests every three months, and request updated documents or revoke access to data.
4. We continue to update our de-identified COVID-19 data set to make COVID-19 data maximally available.
5. We fulfill requests by extracting data from MSDW2 primarily, 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

De-Identified COVID-19 Data Sets

In March 2020, the MSDW team created a de-identified COVID-19 data set containing demographic and clinical information. We built an automated process to refresh this data set weekly and upload it to a secure folder from which end users in the Mount Sinai community may download it via our MSDW website. We also load this data set to the Minerva Data Ark where it is accessible to Minerva users. Our MSDW IRB protocol (STUDY 12-00133) describes our method for de-identifying these data sets.

This data set comprises over 400 data elements in multiple data files. The main data file is the Patient Encounter file, which includes demographic and encounter-level information from our Epic EHR for each COVID-related encounter at the Mount Sinai Health System. The other files -- daily vital signs, medication administrations, lab results, diagnoses, and radiology impressions – can be linked to the Patient Encounter file, as well as to each another, using the masked MRN and masked encounter ID. Through the end of August

2022, more than 300 unique Mount Sinai users have downloaded this de-identified COVID-19 data set more than 6,600 times.

In addition to this de-identified COVID-19 data set, the MSDW team has provided 36 custom data sets of COVID-19 data since August 2021.

MSDW2 Enhancements

Data Content:

New releases of MSDW2 over the last year have added new data content, including the following:

- Clinical notes and reports
- Patient family history
- Immunizations
- Additional patient demographics data
- Patient allergies
- Patient race & ethnicity categories in Epic mapped to standard CDC race and ethnicity codes
- "Pre-defined" patient cohorts for Bio*Me* BioBank participants, for the Image Research Warehouse 1.0, for the Cancer Institute Biorepository, and Mount Sinai Cancer Patient Cohort

Concept Mapping:

The Scientific Computing team has made significant progress in mapping Epic's master files to standard healthcare codes (e.g., ICD-10, LOINC, RxNorm, SNOMED) for the OMOP Common Data Model. We engaged Clinical Architecture, LLC to perform this mapping work using their best-in-class Symedical software product. Clinical Architecture is mapping the following seven Epic master files:

1. Diagnosis (EDG)
2. Procedures (EAP)
3. Surgical procedures (ORP)
4. Lab components (LRR)
5. Medications (ERX)
6. Immunizations (LIM)
7. Allergies (ELG)

Custom OMOP Data Marts:

Scientific Computing is pleased to offer custom OMOP data marts, which are subsets of our OMOP database defined by specific patient cohort inclusion and exclusion criteria. The MSDW team can create such data marts quickly and easily because they are implemented as filtered database views rather than physical copies of data from MSDW2's OMOP database. This approach speeds up delivery time, avoids duplicating data records unnecessarily, and provides configuration flexibility to researchers.

The MSDW team typically implements a data mart for one or more of the following use cases:

1. Repeated data extractions from MSDW2 over time for a particular patient cohort
2. Research teams who want direct-database access to their patient cohort in the OMOP format
3. Research teams who want a dedicated "data source" in ATLAS for their patient cohort

Honest Broker Service:

Some research teams have both PHI and de-identified OMOP data marts of their patient cohort of interest. In these cases, the MSDW team has the ability to generate project-specific masked identifiers in order to provide the research team with a "crosswalk" file of PHI-to-masked identifiers, if requested. (Under no circumstances does the MSDW team reveal its own PHI-to-masked identifier "crosswalk" to anyone.)

Additionally, the MSDW team can serve in an Honest Broker role to link any de-identified data set (or custom OMOP data mart) with any other de-identified data set if the owners of those de-identified data sets register their "crosswalk" files with the MSDW team. This Honest Broker Service allows research teams to combine several de-identified data sets without using or revealing PHI identifiers.

In 2022, the MSDW team has created a de-identification crosswalk file for the Image Research Warehouse 2.0 (IRW) project.

Cohort Query Tools

Leaf is a drag-and-drop cohort query tool that enables researchers to easily create patient cohorts from MSDW2's de-identified data. 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 collaborating with Nicholas on further developing Mount Sinai's instance of Leaf.

Leaf currently supports the following data domains:

- Conditions (diagnoses) using ICD-10-CM codes
- Visit (encounter) types
- Lab results using LOINC codes
- Medications using ATC codes
- Procedures using CPT codes
- Patient demographics, such as current age, ethnicity, gender, race, and vital status
- Vital signs

The MSDW team is continuing to add data elements to Leaf as we integrate more data into MSDW2.

Leaf enables searches of "pre-defined" patient cohorts from Mount Sinai repositories and can export the resulting patient list. Current patient cohorts available for searching are as follows:

1. BioMe Biobank
2. BioMe Biobank Global Diversity Array (Sema4)
3. BioMe Biobank Global Screening Array (Regeneron)
4. BioMe Biobank Whole Exome Sequencing (Regeneron)
5. Cancer Patient Cohort
6. Cancer Institute Biorepository
7. Imaging Research Warehouse 1.0

ATLAS is web-based tool for querying OMOP databases, developed as an open-source software project by the OHDSI community. Users can define their own patient cohorts, browse standard healthcare vocabularies, and conduct cohort-level analysis.

Any Mount Sinai researcher can use ATLAS to query MSDW2's de-identified data without an IRB protocol. The MSDW team also offers researchers the ability to query their custom OMOP data marts of PHI data with an approved IRB protocol.

Data Sources

Image Research Warehouse (IRW) 2.0

This year the MSDW team joined the BioMedical Engineering & Imaging Institute (BMEII)'s project to implement a new Image Research Warehouse (IRW) 2.0 using the open-source software eXtensible Neuroimaging Archive Toolkit (XNAT), originally developed by the Neuroinformatics Research Group at Washington University in St Louis. This new Image Research Warehouse will enable efficient searching across all DICOM imaging files stored within Mount Sinai's GE Centricity PACS system.

The MSDW and BMEII teams are collaborating on a bi-directional interface between IRW and MSDW. The IRW obtains masked identifiers and Epic EHR data elements from MSDW, and the MSDW obtains imaging study metadata parsed from DICOM files extracted from PACS by IRW.

Currently, both teams have established basic data exchange through this interface. Next steps include expanding the scope and functionality of this interface. The MSDW team plans to integrate these DICOM imaging metadata elements into MSDW2's OMOP database for cohort querying within Leaf.

Data Quality

The MSDW team has implemented measures to improve the quality of the data that is being provided to requestors:

1. Improved logic to categorize encounters as office visits, telehealth visits, and emergency department visits
2. Refined calculation of length of stay (LOS) for inpatient visits to handle edge cases and data anomalies

The MSDW team has also implemented the first version of a "rules engine" within MSDW2 that is used to define and execute automated data-quality checks against the MSDW2 OMOP database. We have built an initial set of data-quality rules that this rule engine is now executing in production. We will continue to expand the number, scope, and coverage of these data-quality rules going forward.

Collaborations

INSIGHT Clinical Research Network

Mount Sinai participates in the INSIGHT Clinical Research Network (previously NYC-CDRN), a PCORI-funded network of seven major NYC hospital systems. INSIGHT CRN centralizes patient data contributed by all sites

into an OMOP Common Data Model database to facilitate regional and national research. In accordance with a Data Use agreement with INSIGHT and an IRB protocol approved by the Biomedical Research Alliance of New York (BRANY), the MSDW team provides quarterly data submissions to INSIGHT CRN and bi-weekly submissions of data relevant to COVID-19 research.

PCORnet-RECOVER

Mount Sinai is a participating site in the PCORnet-RECOVER (Researching COVID to Enhance Recovery) project. RECOVER includes pediatric and adult protocols, a prospective cohort study, and data science using EHR data. The MSDW team provides clinical data, geocoded patient addresses, and chart review for this project. MSDW team members also participate in the RECOVER project's EHR Informatics workgroup and QA/QC Data Integrity committee.

COMBATCOVID N3C Collaborative

The National COVID Cohort Collaborative (N3C) is a national initiative being led by the NIH's National Center for Advancing Translational Sciences (NCATS). The goal of N3C is to build a centralized repository of data on patients who are COVID-19 positive or who are COVID-19 symptomatic but untested. The MSDW team contributed data to this initiative in accordance with the N3C agreement.

Sema4

The MSDW team provides data extracts to Sema4 under two different contractual agreements:

1. De-identified data sets for specific research projects under the Data Structuring & Curation Agreement (DSCA)
2. Monthly OMOP-formatted data extracts for all Mount Sinai patients with a cancer diagnosis under the Joint Clinical Annotation Project (JCAP)

The MSDW team implemented the JCAP patient cohort as a custom OMOP data mart for efficiently producing these monthly extract files.

Staffing Update

Over the past year, we have added 5 new members to the team. These positions include a database administrator, 2 database analysts, a computational scientist, and an HPC system administrator. There are currently 14 open positions on the team.

Metrics

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

Year 2022	Electronic Data Capture Tools		Mount Sinai Data Warehouse Query Tools			Custom Data Requests	High Performance Computing
Metric	**eRAP**	**REDCap**	**TriNetX**	**Leaf**	**ATLAS**	**MSDW**	**Minerva**
# User tickets created	794	1,827	4	99	23	261	1,676
# User tickets closed	706	1,649	3	76	16	203	1,649
# Unique users over all time	3,436	17,575	32	63	70	-	3,326
# Active unique logins (last six months)	1,063	5,117	8	52	68	-	762
# User queries or MSDW custom queries	-	-	496	-	-	152	-
# Projects/databases/queries over all time	136	12,780	866	588	-	-	393
# Active projects/databases/queries (last 6 months)	50	4,274	-	499	32	-	337

Goals for the upcoming year

Our roadmap for 2023 includes the following initiatives:

1. Establish a cross-reference between patient records in MSDW and their associated Bio*Me* genetic sequencing files on Minerva.
2. Load genetic variant data from Bio*Me* and other sources to MSDW to enable self-service cohort building in Leaf.
3. Store digitized pathology slide images on Minerva and load their associated metadata to MSDW.
4. Add new data to MSDW including radiology reports; additional Epic flowsheets rows; a cross-walk of historical MRNs for merged patient records; and the admission, discharge, and transfer (ADT) transactions from Epic.
5. Complete the ingestion of DICOM metadata elements from the Image Research Warehouse (IRW) into MSDW for self-service cohort building in Leaf.

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.