MOUNT SINAI DATA WAREHOUSE ANNUAL REPORT

AUG 1, 2023

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, 2022 through August, 2023.

This year we continued to focus 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 provide data to the Mount Sinai user community and to external partners, including Insight CRN (formerly CDRN), National COVID Cohort Collaborative (N3C) and GeneDx (formerly Sema4), while improving our processes and procedures. As of May 2023, we are no longer providing oncology data to GeneDx.

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

1. We only share data sets with requesters who are listed on the IRB approved study (rather than try to collect training information and/or training verification from PIs).
2. We require IRB protocol and approval letter 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 fulfill requests by extracting data from MSDW2 primarily, then from Epic Caboodle or Clarity when necessary, and from other data sources as a fallback. We developed this order of precedence to maximize data quality, timeliness, traceability, and labor efficiency.
5. We have refined the process for researchers and clinicians to request direct database access to clinical data systems and data, such as MSDW2, Epic Clarity and Caboodle or other ancillary systems. All researchers or clinicians who wish to have database access to clinical systems must work with Scientific Computing MSDW team to determine appropriate system and access level. Access to a data mart or other database containing protected Health Information (PHI) within MSDW requires an approved IRB letter, study protocol specifying the inclusion/exclusion criteria and the required data elements, and a signed Database Access Agreement. Approval documents ensure researcher compliance with applicable federal and state laws, and with Mount Sinai's applicable policies for data security, privacy, and compliance. Access to PHI system will be granted only for the duration approved by the study IRB. For access to Epic or other ancillary PHI systems, MSDW will forward the request to the Clinical Data Systems Exception Committee. MSDW participates on this committee

Accomplishments: Aug 1, 2022 – Jun 31, 2023

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)

We currently have mapped 92% of the Epic EHR codes. This will be an ongoing process as new OMOP concept codes are defined.

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

Dashboards

Scientific Computing can now develop custom interactive dashboards using Tableau or Power Bi to visualize essential data and metrics from a pre-defined data set in real time, offering rapid insights into trends and growth. Depending on the request, the dashboards integrate data from multiple data sources including MSDW, Clarity, Caboodle, REDCap, eRAP, the Mount Sinai Cancer Registry and others. Data within a dashboard may be identified or de-identified in accordance with the IRB-approved protocol or the department QI board approval letter.

In the last year, four dashboards have been developed to support research

- Community Outreach and Engagement (COE) Cancer dashboard
- Post Surgery Survivorship dashboard
- Comprehensive Research Data Integration Trust (CReDIT) dashboard
- Alzheimer’s Disease Research Center (ADRC) data core dashboard

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

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** – participants with specimens in the biobank and, optionally, genetic sequencing results in the BioMe database.
    a. **BioMe Biobank Global Diversity Array (Sema4)** – the subset of participants with GDA sequencing results in the BioMe database.
    b. **BioMe Biobank Global Screening Array (Regeneron)** – the subset of participants with GSA sequencing results in the BioMe database.
    c. **BioMe Biobank Whole Exome Sequencing (Regeneron)** – the subset of participants with WES sequencing results in the BioMe database.

2. **Image Research Warehouse 1.0** – the imaging research data warehouse developed by the BioMedical Engineering and Imaging Institute (BMEII), which contains ~525K patient image studies dating back to 2017. Images for patients identified from this cohort may be requested from the BMEII team.

3. **Cancer Institute Biorepository** – participants with specimens (tumor tissue and fluids) from Mount Sinai-affiliated hospitals in the Mount Sinai Cancer Institute Biorepository (MSCIB).

4. **Cancer Patient Cohort** – a "pre-calculated" cohort comprising all Mount Sinai patients with a diagnosis in Epic that maps to a cancer-related ICD-9, ICD-10, or SNOMED code

Enhancements to Leaf functionality in the past year include

- New timelines functionality that allows users to explore the temporal relationship of additional clinical events (aka concepts) to a defined patient cohort
- Improved lab results and measurement concepts that allows users to define specific values or ranges (i.e. equal to, greater than, less than, etc) for lab results or measurements in their query

Data Sources

Mount Sinai's Epic EHR data, sourced from Epic Caboodle, comprises the core set of data within the MSDW2. Data from other databases, systems, and other sources are added to this core set of Epic EHR data to achieve broader and richer coverage. As of July 2023, the other sources that Scientific Computing is using to augment the Epic EHR data set in MSDW2 are as follows:

- MSX billing data warehouse
- Clinisys PowerPath (formerly Sunquest)
- XNAT software that comprises the Imaging Research Warehouse (IRW 2.0)
- CNExT Cancer Registry
- Provation endoscopy system
- Cardiac catheterization lab system
- Social Security Death Master File (DMF)
- Data received from GeneDx (formerly Sema4) under various partnership agreements

The MSDW2 is currently integrating data from PowerPath and the Imaging Research Warehouse's XNAT software into the OMOP data model. The Scientific Computing team anticipates that the data coverage of MSDW2 will expand as Mount Sinai gradually replaces its myriad ancillary systems with their corresponding Epic modules.

Data Quality

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

1. Added all encounter records from Epic in the OMOP visit_occurence table even the "chart documentation events" that aren't face-to-face or telehealth patient encounters
2. Corrected the blood pressure vital signs linkage
3. Improved mapping for race and ethnicity
4. Improved mapping of COVID-19 lab tests

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.

Data Ark Data Commons

The Data Ark Data Commons repository hosted on Minerva HPC consists of public and Sinai generated data sets which are available to researchers with appropriate approvals. There are currently 11 public data sets available, 6 Mount Sinai generated restricted data sets, UK Biobank (restricted), and IBM MarketScan (restricted). There are no restrictions on public data sets. For restricted data sets, requesters must sign data use agreement and provide necessary approvals. Plans for coming year is to provide BioMe Biobank de-identified data sets

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 monthly 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 contributes data to this initiative in accordance with the N3C agreement. Next planned refresh of data to N3C will be in October 2023.

GeneDx (formerly 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).
    a. MSDW received identified and de-identified curated data from GeneDx
2. MSDW provided primary perinatal data to GeneDx and received curated data in return Monthly OMOP-formatted data extracts for all Mount Sinai patients with a cancer diagnosis under the Joint Clinical Annotation Project (JCAP).
    a. We provided monthly data to GeneDx up to Jan 2023
    b. We received identified and de-identified curated datasets from GeneDx in 2022 under this agreement.
    c. As of May 2023, this agreement has been terminated.

Training

In support of ISMMS fellowship program, we have provided EHR data to 3 research fellows in their studies covering Cardiology, Predictors of response to immunotherapy, and Outcomes and Quality of Care in Breast and Gynecologic Cancer Patients. We provided support for 2 TL1/KL2, 3 MD/MSCR, 9 K23/K08/T32 trainees and 44 students in the last year. Our Physician Informaticist provides guidance and clinical research support to provisioning of appropriate data elements and patient cohorts

We held training sessions twice-a-year for all our informatics tools.

Training	**# sessions**
RedCap Introduction	1
REDCap Intermediate	2
Cohort Selection Tools (Leaf/Atlas)	2
Load Sharing Facility (LSF) Job Schedule	2
Containerized Environment on Minerva	1
Running Jupyter Notebook and RStudio on Minerva	1

All training recordings and slides are made available online.

We also held weekly Digital Concierge open office sessions to answer questions about resources and support for researchers. There has been a total of 284 attendees to these sessions for this reporting period.

User Engagement

To engage with and gather valuable feedback from the research community, we hold training sessions on our applications, semi-annual town halls and advisory board meetings, as well as user surveys. Surveys are issued each year on all our applications and responses are collated and posted on our website. In addition, this year we ran a survey and interviewed key stakeholders on the data science needs of the community. Survey questions covered the need for access to expert help, new data sources, informatics tools, and training and workshops. We obtained 146 responses which will be used to inform on prioritization of future initiatives.

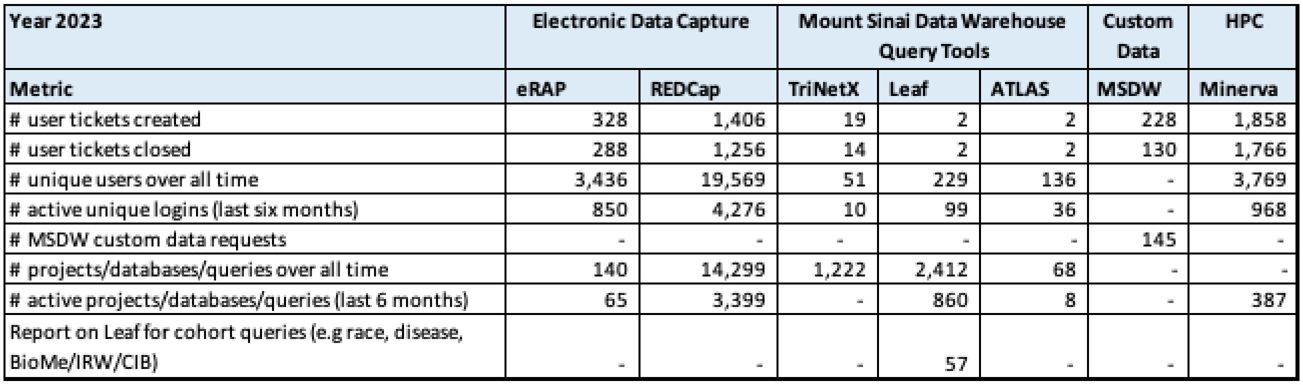
Scientific Computing and Data website provides documentation, webinars, training, and information on our services. All our web site pages contain instructions to researchers on language to acknowledge CTSA in their publications and presentations.

We monitor and track usage of our tools and collect metrics on the number of user tickets submitted to our JIRA ticket system. These are used to improve processes and identify gaps

Metrics

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

Metrics for Year 2023



Year 2023	Electronic Data Capture		Mount Sinai Data Warehouse Query Tools			Custom Data	HPC
Metric	**eRAP**	**REDCap**	**TriNetX**	**Leaf**	**ATLAS**	**MSDW**	**Minerva**
# user tickets created	328	1,406	19	2	2	228	1,858
# user tickets closed	288	1,256	14	2	2	130	1,766
# unique users over all time	3,436	19,569	51	229	136	-	3,769
# active unique logins (last six months)	850	4,276	10	99	36	-	968
# MSDW custom data requests	-	-	-	-	-	145	-
# projects/databases/queries over all time	140	14,299	1,222	2,412	68	-	-
# active projects/databases/queries (last 6 months)	65	3,399	-	860	8	-	387
Report on Leaf for cohort queries (e.g race, disease, BioMe/IRW/CIB)	-	-	-	57	-	-	-

Goals for the upcoming year

Our roadmap for 2023-2024 includes the following initiatives:

1. Integrate Pathology data from PowerPath system into MSDW2 and link to de-identified digitized pathology slides on Minerva
2. 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.
3. 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**.**