# Table of Contents

Introduction .......................................................................................................................... 2
COVID-19 De-Identified Data Sets ......................................................................................... 2
Goal for the upcoming year .................................................................................................... 3
Notable Accomplishments ...................................................................................................... 4
  MSDW data available in new OMOP CDM format .............................................................. 4
  MSDW Hardware Upgrade ................................................................................................... 4
New Data Sources .................................................................................................................. 4
  Prism ...................................................................................................................................... 5
  Provation MD ........................................................................................................................ 5
Increased staffing .................................................................................................................... 5
  Clinical Informaticist – Sharon Nirenberg, MD ................................................................. 5
  Senior Research Data Specialist – Timothy Quinn, PhD ...................................................... 6
Metrics .................................................................................................................................... 6
Data Quality .......................................................................................................................... 6
Collaborations ......................................................................................................................... 7
  CDRN .................................................................................................................................... 7
  Sema4 .................................................................................................................................... 7
Conclusion ............................................................................................................................... 7
Appendix A: MSDW Identifiable Data Attestation Form ....................................................... 8
Introduction

We are pleased to present this annual update on the Mount Sinai Data Warehouse (MSDW) covering the period September 1, 2019 through August 31, 2020. We had a busy year, working to provide data to the Sinai user community, CDRN and Sema4 while improving our processes and procedures. At a high level, we changed the following 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. Developed and now require a form for personnel access for operational data requests (Development office, Press Ganey surveys) that must be updated every three months (Appendix A).
3. Require IRB protocol and approval forms as part of our automated ticket intake process in Jira.
4. Review the dates on IRB, quality improvement, data mart, reports and operational data request every three months and request updated documents or shut off access to data.
5. Developed a de-identified COVID-19 data set to make COVID-19 data maximally available.
6. Follow a “Caboodle-first” policy, where we fulfill requests from Caboodle first, and then Clarity, before querying MSDW. Caboodle data has higher data quality than what is in MSDW, and it is more recent.

COVID-19 De-Identified Data Sets

In an effort to make COVID data available to the Mount Sinai research community quickly as the pandemic surged in New York City at the end of March 2020, the MSDW team created a de-identified COVID data set containing both demographic and clinical information. The data set is refreshed daily and the updated data files are stored in a secure folder for end users in the Mount Sinai community to download via the MSDW website. 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 patient encounter. 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 elements. Furthermore, the data set is enhanced on an ongoing basis to meet the changing needs for analysis as our understanding of the virus and disease improved and has already seen 47 such updates since the start of the global pandemic through the end of this reporting period.

We elected to utilize Epic’s Clarity and Caboodle reporting database as our primary data sources for our de-identified COVID-19 data set for two main reasons. First, we needed to align with the high degree of collaboration established by the interdepartmental COVID-19 Reporting Committee, especially ensuring common data definitions and logic for data extractions from Epic. Second, it was critically important for us to build and iterate on the content of our de-identified COVID-19 data set quickly. We judged that these two aims were best achieved by reducing the number of data-processing steps between the Epic EHR as the source and our data set.

The following graph, Figure 1, shows the number of times the files in the dataset have been downloaded, along with the number of unique users downloading each day. Please note that this graph does not include...
the 683 downloads between March 30, 2020 and April 8, 2020 as they were handled via a different delivery system that did not provide these metrics.

Figure 1: MSDW COVID-19 De-Identified Dataset Downloads

Goal for the upcoming year
Our roadmap for the next year includes building a new OMOP Microsoft SQL Server database fed directly from Caboodle. This first version will cover the core set of data: patient demographics, patient social and family 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.
Notable Accomplishments

MSDW data available in new OMOP CDM format

The Observational Medical Outcome Partnership (OMOP) and Observational Health Data Sciences and Informatics (OHDSI) is a worldwide network of institutions committed to unifying healthcare data from the various systems. In this cohort of over 500 million patients, data is standardized to the OMOP common data model. The OMOP Common Data Model (CDM) is able to assess and analyze multiple data sources concurrently using a common data standard. The common data model stores verbatim data from the source across various clinical domains, such as records for conditions, drugs, procedures, and measurements.

This past year we went live with a conversion of MSDW data into the OHDSI-standard OMOP CDM. This generated both an identified and a de-identified version of a large portion (over 4 million patients) of our dataset for others to use. The de-identified dataset is now available to all ISMMS users via the self-service Atlas platform. Direct SQL access to the identified dataset was also granted to a limited number of researches, per our SOP and relevant IRB protocols and regulations.

MSDW Hardware Upgrade

The MSDW databases and associate processing was primarily running on very old IBM hardware dating back to 2010. These hardware systems approached their vendor supported end-of-life in 2019 and 2020. This hardware supported five logical systems which included the main MSDW production database, the production data transformation processing system, and associated development and quality assurance (QA) systems.

In 2019, the main MSDW production database system was replaced with two new systems in order to:

1. Replace the outdated IBM AIX end-of-life hardware.
2. Improve performance with updated faster CPUs and solid-state drives (SSD) and a duplicate second system to allow for concurrent data updates and analysis.
3. Improve overall MSDW system availability by having a duplicate second system.

During this transition, the Oracle database software was also upgraded since the current version was also near vendor supported end-of-life.

The NetApp Storage Area Network (SAN) disk utilized for the old IBM systems also reached its vendor supported end-of-life and, additionally, was reaching its storage capacity. In Q2 2020, it was replaced by new NetApp Storage Area Network (SAN) hardware. All existing disk storage was migrated to the new storage hardware.

New Data Sources

We do our best to liaison with the relevant stakeholders in our role as an honest broker to obtain data from non-Epic sources. Below are the two examples of how we performed this vital function in the past year.
Prism
With the final go-live of the Epic EHR at Mount Sinai Beth Israel on August 9, 2020, we have a more complete clinical picture of our patients that receive care at our acute-care hospitals. However, for research purposes, it is often important to go back through a patient’s course of treatment to obtain an accurate picture of the services they’ve received in order to properly assign them to a cohort. Prior to their migration to the MSHS Epic platform, Prism was the clinical EHR for the four former Continuum Health Partners Hospitals:

1. Mount Sinai West (formerly Roosevelt Hospital)
2. Mount Sinai Morningside (formerly St. Lukes Hospital)
3. Mount Sinai Beth Israel (formerly Beth Israel Medical Center)
4. Mount Sinai Brooklyn (formerly Beth Israel Kings Highway Division)

Now that Prism is no longer in production use, we have taken the opportunity to ensure that we have continued access to the database in a read-only mode. This allows us to be able to provide a detailed clinical picture of our patients, straddling the various transition points between the Prism and Epic EHRs. Using this access has enabled us to provide data on two requests since the shift over to the Epic system.

Provation MD
Over the course of the past year, MSDW has received a handful of requests for more detailed information about Gastroenterology procedures that are documented in the Provation MD system. Although an interface exists between this system and Epic, only a portion of the data associated with each documented procedure are transferred to Epic. Furthermore, even the data that is ultimately stored in Epic is recorded in a free-form manner – inhibiting its discovery and retrieval. After a number of months of concerted effort, with the assistance of the department and MSHS IT staff, we were able to obtain a copy of the Oracle database underlying the application and secure daily updates. This gives us the ability to see the raw discrete data, allowing us to fulfill the requests that specifically target the information available in this database and provide it to the MSHS community. In the most recent two months, since we acquired this access, we have already provided data from the Provation MD for two data requests.

Increased staffing
Over the past year, we have strategically grown our team by adding two key senior positions. The first is a full-time clinical informaticist, Sharon Nirenberg, MD. Her role will give us dedicated time from an extremely well-qualified clinical informatics professional, which we hope will improve our communication and collaboration with researchers and other data requestors. The second is Timothy Quinn, PhD, who joins us as a Senior Research Data Specialist. His role will to be to improve our environment in ways that help our team become more efficient in fulfilling our mission of delivering data to the ISMMS research community. Please see a short biography of each of them below and join us in welcoming them to the ISMMS family.

Clinical Informaticist – Sharon Nirenberg, MD
Sharon received her medical degree from Rutgers University New Jersey Medical School. Subsequently she obtained a Master’s of Science in Medical Informatics from Northwestern University and went on to complete a Clinical informatics fellowship program at Northwell Health. Upon finishing the fellowship, she served as a Senior Specialist, Clinical Informatics in the Ambulatory Informatics group at Northwell Health.
this role she collaborated with clinical leads to build and design specialty specific dashboards embedded within the Electronic Health Record. The application contained logic driven, passive clinical decision support.

Senior Research Data Specialist – Timothy Quinn, PhD
Timothy has over 10 years of experience designing and leading the implementation of data warehouses and business intelligence reporting in the healthcare industry. He is an expert in the methods for integrating data from electronic health records (EHRs), ancillary systems, healthcare claims, and public data sets, including master and reference data management, person matching, and record versioning. He has designed custom healthcare data models and has implemented the major healthcare data models from Oracle, IBM, Epic Systems Corporation, and is now working with OHDSI’s Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) upon joining Mount Sinai. He has also worked as an industrial engineer for several academic medical centers, including Vanderbilt University Medical Center and the University of Texas MD Anderson Cancer Center, focusing on process analysis and redesign, performance measurement, data analysis, and Lean Six Sigma quality improvement projects. Dr. Quinn holds Bachelor’s and Master’s degrees from Stanford University’s School of Engineering, and a Ph.D. in System Dynamics from the Massachusetts Institute of Technology.

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>I2B2</th>
<th>TriNetX</th>
<th>MSDW</th>
<th>CQT</th>
<th>OMOP</th>
</tr>
</thead>
<tbody>
<tr>
<td># of user tickets created</td>
<td>40</td>
<td>12</td>
<td>634</td>
<td>159</td>
<td>4</td>
</tr>
<tr>
<td># of user tickets closed</td>
<td>37</td>
<td>7</td>
<td>839</td>
<td>214</td>
<td>3</td>
</tr>
<tr>
<td># of unique users</td>
<td>186</td>
<td>52</td>
<td>-</td>
<td>859</td>
<td>62</td>
</tr>
<tr>
<td># of active unique users</td>
<td>37</td>
<td>49</td>
<td>-</td>
<td>176</td>
<td>-</td>
</tr>
<tr>
<td># queries run by users</td>
<td>722</td>
<td>306</td>
<td>219</td>
<td>7,194</td>
<td>0/6</td>
</tr>
<tr>
<td># active projects /data marts (last 6 months)</td>
<td>-</td>
<td>-</td>
<td>20</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td># total projects/data marts</td>
<td>-</td>
<td>-</td>
<td>20</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td># collaborations/clinical trials started</td>
<td>-</td>
<td>10/73</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Data Quality
The MSDW team has put in place several measures to improve the quality of the data that is being provided to requestors. Arguably, the most critical change has been to pull data directly from the Epic Clarity and Epic Caboodle databases for the majority of data requests, instead of the MSDW database. Pulling the data directly from Epic databases whenever possible, allows the team to provide cleaner data, as the data is coming from one source system instead of multiple systems as is the case with MSDW. Data from the MSDW database is now used to supplement data provided from Clarity and Caboodle whenever necessary.
Collaborations

CDRN
Mount Sinai participates in INSIGHT Clinical Research Network (previously NYC-CDRN), a PCORI funded network of seven major NYC hospital systems, that 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 uploads to the INSIGHT and since March 2020, bi-weekly uploads of data relevant to COVID-19 research.

Sema4
We regularly provide de-identified or identified datasets to Sema4 through specific research projects as well as through legal agreements such as the DSCA.

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.

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.
Appendix A: MSDW Identifiable Data Attestation Form

Mount Sinai Data Warehouse - Identifiable Data Attestation Form

*For Hospital Operations ONLY*

As per compliance at the Mount Sinai Health System, the heads of departments requesting protected health information (PHI) for hospital operations initiatives, must identify project team members who can receive data sets containing PHI. This list must be updated by the department head every 4 months. Data sets containing PHI will ONLY be released to project team members listed on the unexpired attestation form below.

Project name: _________________________________________________________________

Project Description:______________________________________________________________

_____________________________________________________________________________

Name and title of project team members who are eligible to receive identifiable data:

_____________________________________________________________________________

_____________________________________________________________________________

_____________________________________________________________________________

_____________________________________________________________________________

_____________________________________________________________________________

Signature below attests that the current project team members listed above are eligible to receive identifiable data for this project.

Department:_______________________________________________________________

Department Head Name:______________________________________________________

Department Head Signature:__________________________________________________

Date Signed:_________________________________________________________________

Expiration Date (4 months from Date Signed):____________________________________

*Please attach this completed form to your MSDW data request ticket.*

Page 8 of 8