2023 Clinical and Translational Science (CTSA) Data Science Survey: Responses

Introduction

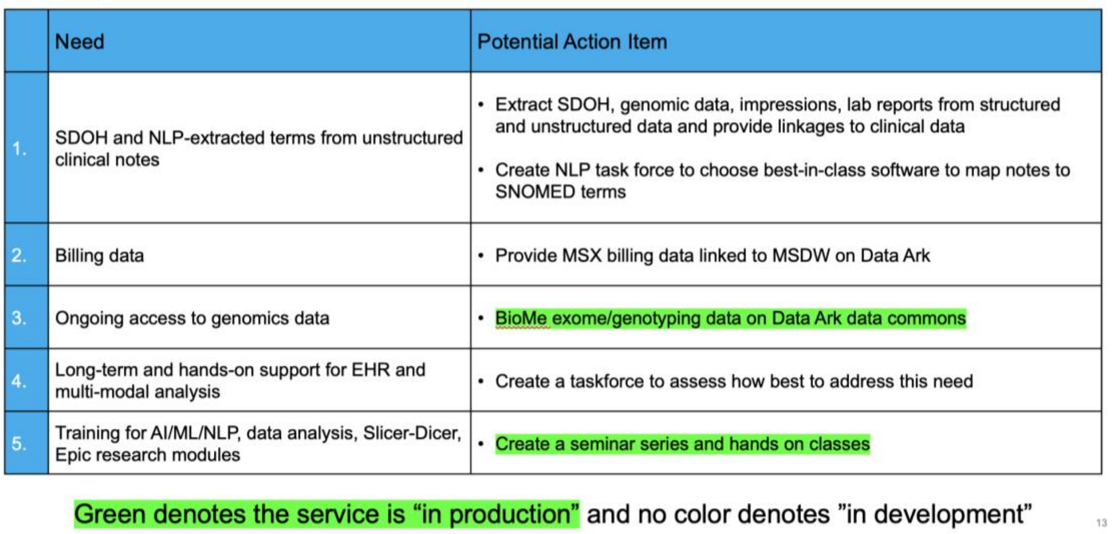
The purpose of this survey, conducted in February 2023, was to gather feedback from Mount Sinai research community on the barriers to leveraging data science in their research.

Two main components included an electronic survey distributed via email with 146 responses; 1:1 interviews with 12 key thought leaders were also held.

The four main topics covered were:

- Access to expert help
- New data sources
- Access to Informatics tools
- Training/workshops and seminars

A summary of highest-priority needs identified by the survey, with potential action items, is provided in the table below.



	Need	Potential Action Item
1.	SDOH and NLP-extracted terms from unstructured clinical notes	• Extract SDOH, genomic data, impressions, lab reports from structured and unstructured data and provide linkages to clinical data • Create NLP task force to choose best-in-class software to map notes to SNOMED terms
2.	Billing data	• Provide MSX billing data linked to MSDW on Data Ark
3.	Ongoing access to genomics data	• BioMe exome/genotyping data on Data Ark data commons
4.	Long-term and hands-on support for EHR and multi-modal analysis	• Create a taskforce to assess how best to address this need
5.	Training for AI/ML/NLP, data analysis, Slicer-Dicer, Epic research modules	• Create a seminar series and hands on classes

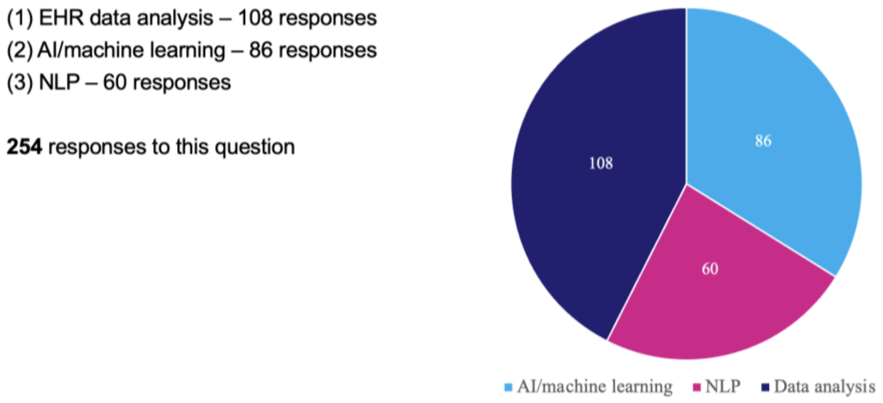
Green denotes the service is “in production” and no color denotes ”in development”

Survey questions:

What will help you overcome the barriers to leverage data science in your research? (Choose one or more)

1. **Access to experts specializing in**
   - AI/Machine Learning
   - Natural Language Processing (NLP)
   - EHR Data analysis

Responses:

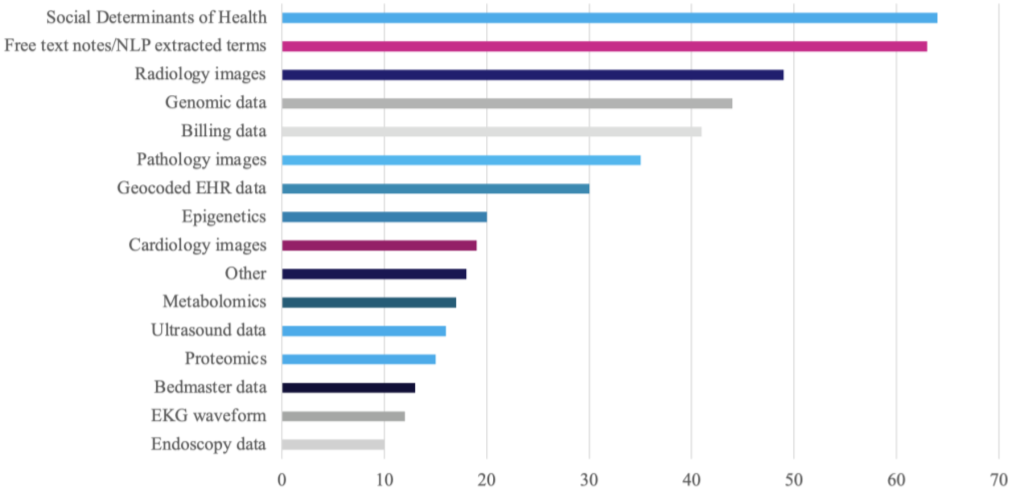


(1) EHR data analysis – 108 responses
(2) AI/machine learning – 86 responses
(3) NLP – 60 responses

254 responses to this question

2. **Access to data resources specializing in**
   - Free text notes for provider, pathology reports, etc.
   - Geocoded EHR data
   - Radiology images
   - Pathology images
   - Cardiology images
   - Bedmaster data
   - Billing data
   - Genomic data
   - EKG waveform
   - Ultrasound data
   - Social Determinants of Health
   - Endoscopy data
   - Epigenetics
   - Proteomics
   - Metabolomics
   - Other: Please specify new data sources

Responses:

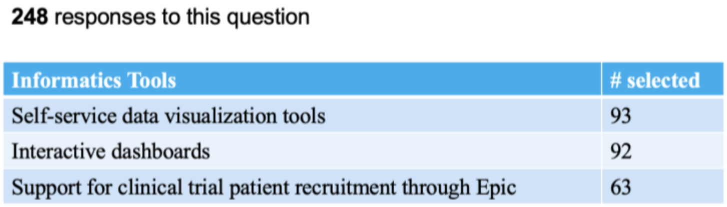


(X-axis: Number of responses)

3. Access to informatics tools and services

- Service to create an interactive dashboard for your research
- Access to self-service data visualization tools
- Support for clinical trial patient recruitment through Epic

Responses:



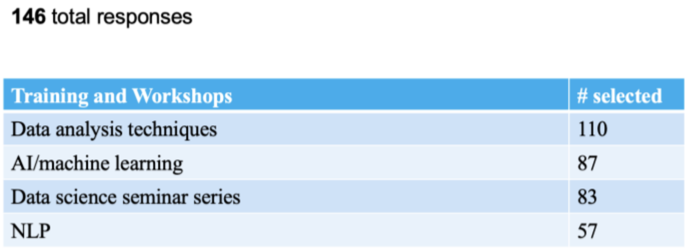
248 responses to this question

Informatics Tools	# selected
Self-service data visualization tools	93
Interactive dashboards	92
Support for clinical trial patient recruitment through Epic	63

4. Access to training/workshops and seminars

- NLP
- AI/machine learning
- Data analysis techniques
- Data science seminar series

Responses:



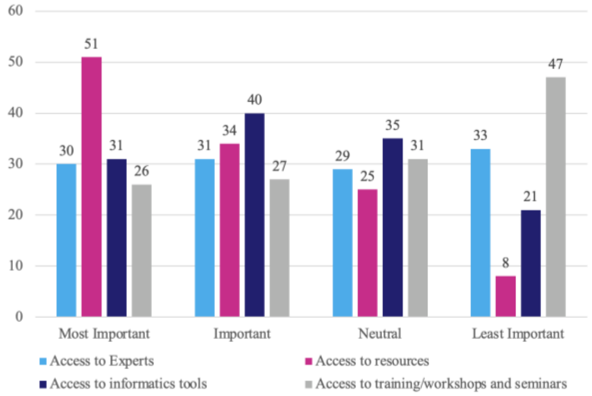
146 total responses

Training and Workshops	# selected
Data analysis techniques	110
AI/machine learning	87
Data science seminar series	83
NLP	57

5. Provide a ranking from most important to least important

- Access to experts
- Access to resources
- Access to informatics tools
- Access to training/workshops and seminars

Responses:



- ➢ Access to Resources was ranked most important
- ➢ Access to Long term expert help/ Informatics Tools ranked next important
- ➢ Access to training/workshops was ranked least important overall

(Y-axis: Number of responses)

Submitted Comments with Responses

This section covers the open comments section of the survey, and responses.

Additional Data Sources

Survey Responder Comment	Response
Acquire and link public data sets to Mount Sinai's Epic EHR data in MSDW. Examples: U.S. Census Bureau's American Community Survey, the CDC's NHANES, other data sets on data.gov	The results of the 2022 United States Census Bureau's American Community Survey (ACS) are now stored in the Mount Sinai Data Warehouse (MSDW). Additionally, all current and historic patient addresses from Epic are now geocoded using the Decentralized Geomarker Assessment for Multi-Site Studies (DeGAUSS) application and are also stored in MSDW. Using the geocoded patient home addresses, the MSDW team can readily link the results of the ACS to the patients in Epic.

Survey Responder Comment	**Response**
other OMICs data clinical, serological, immunological (which may already be available through Epic)	The Mount Sinai Data Warehouse (MSDW) contains all the clinical, serological, and immunological data that is available in Epic. For data not available in Epic, please submit a request with the MSDW team and the team will evaluate if it is feasible to obtain access to this data from the data owner and store this additional data in the MSDW.
Labs, lab trajectories, body weight trajectories, height vaccination and infection history for differing pathogens data on medication usage Nursing data and quality metrics It would be great to have clinical data that includes cytokine, chemokine, and lymphocyte concentrations in serum or CSF Access to psychiatric notes Patients report of pain or any information related to pain perception by provider or patient Clinical data -- lab and ICD coding Culture Results (microbiology) Radiology notes EMRs for symptoms or diagnoses, patient names/phone numbers	The Mount Sinai Data Warehouse contains patient level demographic information, lab results, culture results, diagnoses (including the associated ICD-10 CM code), vital sign measurements (blood pressure, oxygen saturation, pain scale), immunization orders and administrations, medication orders and administrations, microbiology results, and nursing flowsheets that are recorded in Epic as structured data. The MSDW also has access to all the free text clinical notes recorded in Epic, including imaging and pathology reports. Free text clinical notes often contain more detailed symptomology information. Please see the following website for detailed information about the data stored in the MSDW https://labs.icahn.mssm.edu/msdw/data-sources/
Would be interested in getting access to billing data, reports (radiology, endoscopy, pathology) and endoscopy data to complement the data that is present in MSDW2	The Mount Sinai Data Warehouse team has access to data sources, that are not yet integrated into OMOP. Billing data from the MSX database, radiology and pathology reports in Epic, and endoscopy data from Provation can be provided in a custom data set curated by the MSDW group.

Survey Responder Comment	Response
It would help to have a useful GUI interface and the ability to download data for analysis in our stats program of choice	For all custom data requests, data is delivered in a pipe de-limited text file to facilitate uploading the data to any statistical program of choice. The Mount Sinai Data Warehouse contains all lab results for serum and CSF that are recorded in Epic as structured data. The psychiatry notes are not available in Epic.
I am looking for language datasets related to neuropathic pain Oncology outcome data (OS, RFS, RECIST) data sets on exposure or usage of opioids	The Scientific Computing and Data Division is always open to providing researchers access to additional data sets. Please submity a request with the MSDW team, including details about the specific data set being requested, and the team will investigate the feasibility of obtaining access to this data.
I need access to radiology images (xrays for example) and SafetyNet data It is not clear to me how I can access radiology images I need access to echocardiographic images and MRIs images from PACS. I would like more training in deep learning	The Imaging Research Warehouse managed by the BioMedical Engineering and Imaging Institute (BMEII), can provide researchers with radiology images. For echocardiographic images, please contact the Cardiology IT team.

Data Visualization and Dashboards

Survey Responder Comment	Response
I have been asking for visualization and dashboarding tools for years now, we are behind the competition for accelerating research because our IT infrastructure is bottlenecked - limited and the IT group is understaffed to support the bandwidth the researchers need.	The Mount Sinai Data Warehouse team offers a service that develops Tableau dashboards for researchers. Please open a request with the MSDW team for more information.

Either create a steering group that includes data oriented research staff with the IT experts or invest the R dollar gains into universal basic data management and visualization tools	

Data Ark Data Commons

Survey Responder Comment	Response
Synthetic biology datasets. Genotype (genomics) to phenotype (e.g. cell painting images, or others) association datasets. Representation learning with multimodal datasets. Mass Spectrometry based metabolomics and exposomics datasets Access to large databases for research such as the IBM Marketplace is key for early researchers who do not have funds to generate pilot data for career awards and grants	There are currently 19 data sets hosted on the Data Ark Data Commons including both genotype and phenotype data. The IBM MarketScan data is also available on Data Ark. The Data Ark team routinely surveys the Mount Sinai research community for recommendations of high-impact, and shareable data sets to onboard (See https://redcap.link/suggest_data). In addition, to express interest in data hosting on Data Ark, Principal Investigators (PIs) can to outline data specifications and target research groups by submitting this form: https://redcap.link/data_intake

AI/ML

Survey Responder Comment	Response
AI and NLP would keep us competitive with NYU and Northwell Access to the boots-on-the-ground folks who can do the actual data pipeline coding and appropriate statistical analyses	The Scientific Computing and Data Division is investigating AI/ML solutions for the Mount Sinai Data Warehouse.
AI prognostic models for cancer recurrence based on available lab data, pathology, imaging. Ability to screen patients in the EHR by	The Scientific Computing and Data Division agrees, and the team is investigating AI/ML solutions for the Mount Sinai Data Warehouse.

disease type to develop an idea for a project.	In terms of searching the electronic health record data to develop ideas for a project, please use one of the 3 cohort query tools, including Leaf, ATLAS, and TriNetX, that are supported by the Scientific Computing and Data Division. Previous training sessions on these tools can be found on the Mount Sinai Data Warehouse website.
Looking to use AI as a resource to search for language related to pain and pain perception by patients - I would love to have training in how to establish an algorithm that could track this in specific pain related areas of care	In Fall 2023, Dr. Girish Nadkarni and Dr. Hayit Greenspan led a course entitled AI/ML in the Clinic.
The issue is less so about access to data sets and tools but more so about addressing governance roadblocks around linking different data sets together, e.g. genomic data and clinical notes. In the current scenario, the technological workaround is to build a note-deidentification system that the IRB and other review teams at Mount Sinai are comfortable with.	The Scientific Computing and Data Division is researching tools and methodologies for de-identifying the clinical notes that also meet the privacy requirements of the Icahn School of Medicine at Mount Sinai Program for the Protection of Human Subjects (PPHS).
I think that more lectures/seminars on imaging processing/ML techniques would be really useful and more resources (these are important but I couldnt rank them higher!)	Students run workshops occasionally on advanced image processing. One such workshop is called Diffusion Weighted Imaging Processing and Analysis and is led by Mackenzie Langan. Some lectures and seminars on new image processing tools have been recorded and can be provided upon request to the BMEII team.

Cohort Query Tools

Survey Responder Comment	Response
Data accessible through Atlas and Leaf would be much more useful if complemented with clinical outcomes	Please contact the Mount Sinai Data Warehouse team to request additional data elements you would like to see in Leaf and ATLAS.
It's essential that Clinical Research Coordinators are allowed to maintain access to Slicer Dicer and its exports for patient recruitment	Epic SlicerDicer should not be used for research trial recruitment as the protected patients are included in this database that SlicerDicer queries. Please contact the Mount Sinai Data Warehouse team for assistance with clinical trial recruitment.

Training

Survey Responder Comment	Response
Having experts to consult with would be helpful, but resources to the tools with online guides on how to use them would be the most useful.	Links to all trainings and information on our services are posted on the Scientific Computing & Data website There are a number of trainings available on PEAK: REDCap: REDCap Application Training; Leaf cohort query tool: Written Tutorial; PEAK Tutorial Atlas cohort query tool: Written Tutorial; PEAK Tutorial; Videos TriNetX cohort query tool: PEAK Tutorial Information and training on Observational Medical Outcomes Partnership (OMOP) Common Data Model is available on our website Digital Concierge open hours are held each Wednesday where you can speak representatives of our services.

Operations

Survey Responder Comment	Response
Hard to troubleshoot problems past the entry level -MSDW data pulls - RedCAP There's only very high level or entry level help - the process for feedback or improvement is very manual and depends on user push.	The Scientific Computing and Data Division regularly monitors the status of the finance process. If you have a hard deadline that is being threatened by the length of the billing process, please contact the Mount Sinai Data Warehouse team or contact the REDCap team.

There are multiple dashboards that require a steep learning curve but no one to help get past the lowest level. Finance/billing delays data pull - both MSDW and my department grants and finance do not adequately support the data scientists, seem to rely on the requesting team and data scientists to push the steps of billing from one payment step to another instead of escorting it through the end of the process themselves.	

Custom Data Requests

Survey Responder Comment	**Response**
Data requested for IRB approved projects with PIs from departments that have that data should not be charged for these requests.	Charging for custom research data requests is required to sustain the operational costs of the Mount Sinai Data Warehouse.
Don't like the way to above question was structure -it was not a rank-am not neutral-its all important. We desperately need access to data to do feasibility assessments for clinical trials and to screen for trial participants, including more detailed searches than are possible with existing self service tools. Also, given the time it takes to accommodate requests at time for data, we need to be able to request a search in parallel with IRB approval understanding that the results may not be made accessible to use before IRB approval	The Icahn School of Medicine at Mount Sinai Program for the Protection of Human Subjects (PPHS) requires that a study protocol, delineating all requested data elements, be approved by the IRB office before any protected health information is shared with the research staff.

REDCap

Survey Responder Comment	Response
The clinical data pull from Epic is great, but I wish it was more accessible.	The REDCap support team is available to answer questions about Epic Clinical Data Pull through a REDCap support request, as well as during weekly Digital Concierge sessions and scheduled consultations.
Available assistance in general for redcap users	Introductory REDCap training sessions are offered twice a year. Additionally, the REDCap team conducts advanced training sessions. See our REDCap site for training links and announcements. Additionally, the REDCap team attends the Digital Concierge open hours each Wednesday where you can go for REDCap assistance and scheduled consultations can be arranged.
informatic tools to streamline data collection between REDCap databases and other databases	The REDCap support team is available to answer questions about REDCap data collection options and tools through a REDCap support request, as well as during weekly Digital Concierge sessions and scheduled consultations.

High Performance Computing

Survey Responder Comment	Response
How to implement tools that have Docker/Singularity dependencies without space issues -How to run neuroimaging pipelines -more Job submission guidance	The Minerva team does not support Docker on the Minerva High Performance Computing platform for security reason. However, we do support Singularity. Here is our guide on how to use singularity on Minerva HPC: https://labs.icahn.mssm.edu/minervalab/documentation/running-container-singularity/ Here is our guide for how to submit LSF jobs on Minerva: https://labs.icahn.mssm.edu/minervalab/documentation/lsf-job-scheduler/ If you still have issue with singularity and LSF job submission, please email hpchelp@hpc.mssm.edu. We

	will work with you on the submission script for your pipeline.
It is not clear to me how to use the Minerva supercomputer	Regular training sessions are held for Minerva. Please see the Minerva Lab website for past training materials and announcements on future sessions

Epic for Research

Survey Responder Comment	**Response**
training sessions on how to optimize EPIC data for research recruitment.	Epic Research training covering research recruitment topics was offered on Tuesday, November 7, 2003 and will be offered again during 2024.