

# Mount Sinai Data Warehouse Town Hall

Scientific Computing and Data  
Icahn School of Medicine at Mount Sinai  
November 20, 2024



Icahn  
School of  
Medicine at  
**Mount  
Sinai**

# Agenda

1. MSDW Operations
2. MSDW Major Accomplishments
  - Obtaining somatic genomic results from external vendors
  - IRW 2.0 searchable in Leaf
  - Digital Pathology
3. MSDW Roadmap November 2024 – May 2025

# MSDW Operations

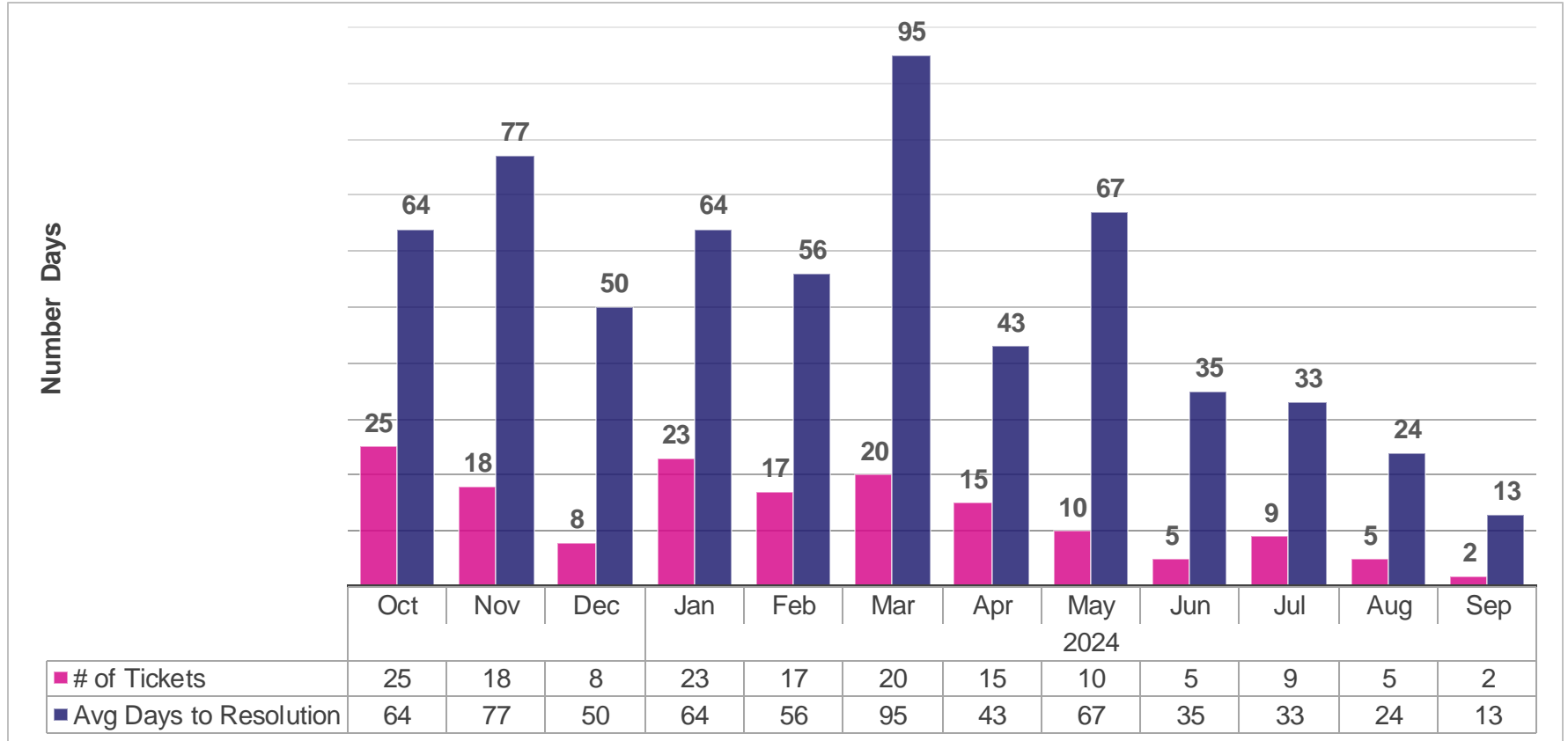
# MSDW Data Sets Delivered

- ▶ From January 2024 to September 2024, the MSDW team closed 106 data requests
  - This represents about a 50% decrease in the number of data requests closed in the same time frame in 2023
- ▶ There has been a 40% decrease in the data set delivery time from January to September 2024 compared to the same time frame in 2023

## Top Department Requesting MSDW Data Sets

Department	Tickets (N)
Medicine	17
Population Health	13
Cardiology	7
Genetics	5
Immunology	4

# Custom Data Set Average Days to Resolution

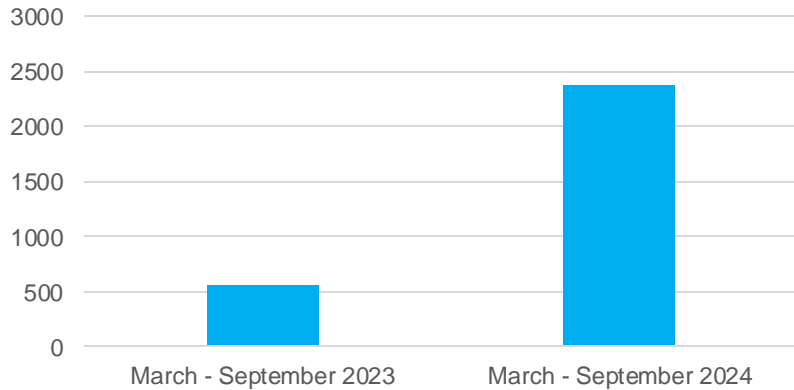


# Utilization of Patient Cohorts and New Features in Leaf

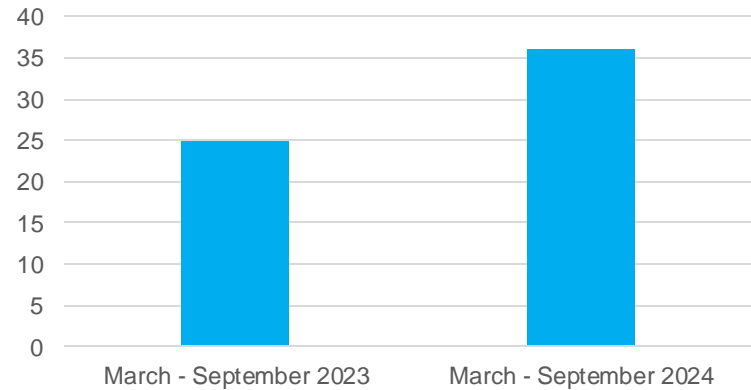
Patient Cohort/ Feature	Distinct Users Executing Queries Patient Cohorts (N)
Cancer Staging from Cancer Registry	17
BioMe BioBank	31
BioMe Biobank Global Diversity Array (Sema4):	9
BioMe Biobank Global Screening Array (Regeneron)	3
BioMe Biobank Whole Exome Sequencing (Regeneron)	9
Cancer Institute Biorepository	9
Cancer Patient Cohort	20
Dental Patient Cohort	1
Digitized Pathology Slides Cohort	2
Imaging Research Warehouse 1.0	7
Imaging Research Warehouse 2.0	5

# TriNetX Utilization Continues to Increase

Comparison of the Number of TriNetX Queries Run in March to September 2024 vs. the same time frame in 2023



Number of New TriNetX Users from March - September 2024 Compared to March to September 2023



**Over 4 times the number of queries were run in TriNetX from March to September 2024 compared to March to September 2023**

# Utilization of Geocoded Patient Addresses

- ▶ **Over 10 million current and historic patient address**
- ▶ **Geocoded patient addresses provided in 4 custom data requests**
- ▶ **Geocoded patient addresses are stored in the Mount Sinai Data Warehouse**
  - Used the Decentralized Geomarker Assessment for Multi-Site Studies (DeGAUSS) application
  - Patient home address recorded in Epic converted to latitude and longitude points
  - Processes conducted on current and historic patient home addresses
- ▶ **The 2022 American Community Survey (ACS) available in MSDW**
  - Yearly survey conducted by the United States Census Bureau
  - ACS results link demographic, social, economic and housing data to geographic points
  - Mechanism to link Social Determinants of Health (SDoH) data to geocoded patient addresses



# Outreach to MSDW Users: March 2024 – November 2024

Date	Event	Participants (n)
October 14, 2024	Epic for Research Training	125
October 9, 2024	TriNetX Training Session	16
October 10, 2024	Leaf and ATLAS Training Session	7
October 17, 2024	Presented to Systemwide Clinical Research Meeting on Return of Foundation Medicine Results	65
Every Wednesday	Digital Concierge	410
	<b>TOTAL</b>	<b>623</b>

# Funding and Publication Supported by MSDW

# of pubs 2023	# of pubs since 2012	# of selected high impact pubs***	# of high impact pubs since 2012	# of citations of all pubs	Amount of funding**
39	146	2	6	3,726	\$24,712,055

\*\* Subcontracts from other organization were only included if reported by the PIs or used for chargebacks

\*\*\* Journals with impact factor >= 30

## Methods used to quantify MSDW supported funding and publication

Publication collection method	PI response rate for pubs	Funding collection method	PI response rate for funding
PI report via survey	61/157	PIs confirm via survey	10/82 PIs confirmed 10 NIH awards
		Paid by NIH awards	13 NIH awards
		Data request record	2 NIH awards from data request records

## Require MSDW Users to Agree to Acknowledge the CTSA

As of January 1, 2024, anyone who requests a custom data set from the Mount Sinai Data Warehouse must agree to cite the CTSA in any publications resulting from the requested data set.

The agreement is part of the ticket intake process.

This is because of the support provided by the CTSA for the MSDW.



Supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences, National Institutes of Health.

# MSDW Major Accomplishments

# Obtaining Somatic Genomic Results from External Vendors

## ▶ **Project objectives**

- To link the phenotypic and somatic genomic data on Data Ark, facilitating the use of somatic genomic data for analytics, clinical research and clinical operations
- To make raw and structured somatic genomic results from external vendors available to the Mount Sinai research and clinical community




## ▶ **Collaboration with Mount Sinai Innovation Partners (MSIP)**

- MSIP ensuring contracts address Mount Sinai's best interest for use of somatic genomic data

## ▶ **Genomic results to be stored on Minerva**

- Results include both structured and raw genomic data
- File formats received include BAM, FASTQ, VCF, PDF, XML, JSON and CSV
  - File types available vary by vendor

# Somatic Genomic Testing Vendors Currently Engaged

Vendor	Status
<p data-bbox="301 312 643 342">Foundation Medicine</p> 	<ul data-bbox="736 312 1700 517" style="list-style-type: none"><li>• Contract signed</li><li>• All historic results stored on Minerva</li><li>• New results received daily and stored on Minerva</li><li>• Raw and structured results available via request to Data Ark and MSDW teams</li></ul>
<p data-bbox="301 550 388 580">Caris</p> 	<ul data-bbox="736 550 1591 672" style="list-style-type: none"><li>• Contract signed</li><li>• New results received daily and stored on Minerva</li><li>• Historic results expected by December 2024</li></ul>
<p data-bbox="301 705 533 736">NeoGenomics</p> 	<ul data-bbox="736 705 1561 823" style="list-style-type: none"><li>• New Lab Service Agreement (LSA) required</li><li>• LSA reviewed by Mount Sinai and now awaiting comments from NeoGenomics</li></ul>

# Data Transferred from Foundation Medicine to Minerva

File Type	# of Orders Resulted	Size
BAM	8,161	~45,000 GB
VCF	7,174	~7 GB
PDF	8,161	~8 GB
JSON	3,090	~3 GB
XML	8,161	~8 GB

- ▶ About the data from Foundation Medicine:
  - Includes results of orders for a Foundation Medicine test placed at a Mount Sinai facility
  - All historic and current Foundation Medicine results
  - Structured and unstructured results
  - Updated daily with new results

# Exposing Pathogenic Gene Mutations in Leaf for Clinical Trial Feasibility Assessment

Genes Identified by Oncology	
ALK (Anaplastic Lymphoma Kinase)	MAGE-A4
ATM (Ataxia Telangiectasia Mutated)	MET
ATR	MSI-H
BRAF (B-Raf Proto-Oncogene, Serine/Threonine Kinase)	MTOR
BRCA1/BRCA2	MYC
CDK4/6	NTRK (Neurotrophic Tropomyosin Receptor Kinase)
CTLA-4 (Cytotoxic T-Lymphocyte Antigen 4)	PD-L1
DDR	PIK3CA
EGFR (Epidermal Growth Factor Receptor)	PTEN
EZH2	RET
FGFR1/2/3 (Fibroblast Growth Factor Receptor)	ROS1
HER2 (ERBB2)	TP53
IDH1/IDH2	TROP2
KRAS	VEGF/VEGFR
LAG-3	Wnt/ $\beta$ -catenin Pathway



# Foundation Medicine Patient Cohort Searchable in Leaf

The screenshot displays the Leaf application interface for searching patient cohorts. The browser address bar shows leaf.mssm.edu. The top navigation bar includes 'Unsaved Query' with '0 patients', '+ New Query', 'Databases', and the user 'sharon.nirenberg'. A sidebar on the left contains navigation options: 'Find Patients', 'Visualize', 'Timelines', and 'Patient List'. The main content area features a search bar with 'All Concepts' and 'Search...' dropdowns, and a green 'Run Query' button. Below the search bar is a 'Limit to' section with three filter panels, each containing 'Patients Who', 'Anytime', and 'At Least 1x' dropdowns. The central list of cohorts includes:

- Cancer Registry
- Conditions (ICD-10-CM)
- Demographics 11,616,181
- Encounters 6,046,450
- Imaging Research Warehouse
- Lab Results & Measurements (LOINC)
- Medications (ATC)
- Patient Cohorts 2,667,759
  - BioMe Biobank 53,918
  - BioMe Biobank Global Diversity Array (Sema4) 9,565
  - BioMe Biobank Global Screening Array (Regeneron) 13,604
  - BioMe Biobank Whole Exome Sequencing (Regeneron) 1,176
  - Cancer Institute Biorepository 13,682
  - Cancer Patient Cohort 235,535
  - Dental Patient Cohort 55,200
  - Digitized Pathology Slides Cohort 1,000,000
  - Imaging Research Warehouse 1.0 242,491
  - Imaging Research Warehouse 2.0 1,111,111
  - Foundation\_Medicine 4,003**
- Procedures (CPT4)
- Vitals 3,315,306
- Medical Subjects

A pink callout box highlights the 'Foundation\_Medicine' cohort entry in the list, showing a person icon, the name 'Foundation\_Medicine', and a count of '4,003'. Another pink box highlights the 'Foundation\_Medicine' entry in the search results area, showing a person icon, the name 'Foundation\_Medicine', and a count of '4,003'.

# Query the Imaging Research Warehouse 2.0 in Leaf

## ► New functionality

- In **Leaf**, users can search for images in the **IRW 2.0** by:
  1. Imaging modality (i.e., Computed Tomography, Digital Radiography, etc.)
  2. Body part/ Procedure
- Identify cohorts of patients with specified clinical characteristics recorded in Epic and have certain types of de-identified images in the IRW 2.0

The screenshot displays the Leaf interface for querying the Imaging Research Warehouse 2.0. The interface is divided into several sections:

- Left Sidebar:** Contains navigation options: Find Patients, Visualize, Timelines, and Patient List.
- Top Bar:** Shows "Unsaved Query" with "0 patients", a "+ New Query" button, "Databases" dropdown, and the user name "sharon.nirenberg".
- Search Area:** A search bar labeled "All Concepts" with a search icon and a search input field.
- Concepts List:** A tree view of concepts. The "Imaging Research Warehouse" category is expanded, showing a list of imaging modalities: BONE DENSITY SCAN/ DIGITAL RADIOGRAPHY, COMPUTED TOMOGRAPHY (CT), DIGITAL RADIOGRAPHY, MAGNETIC RESONANCE, MAMMOGRAPHY, NUCLEAR MEDICINE, OTHER (IMAGING MODALITY), POSITRON EMISSION TOMOGRAPHY (PET), RADIO FLUOROSCOPY, and ULTRASOUND.
- Right Panel:** A detailed view of the selected "Imaging Research Warehouse" category, listing the same imaging modalities as the left panel. It includes a "Bin Query" button and a filter section with "And" and "Anytime" options, and a "At Least 1x" filter.

# Query the Imaging Research Warehouse in Leaf

The screenshot shows the Leaf interface for querying the Imaging Research Warehouse. At the top, a blue header displays the Leaf logo, a search bar, and a query summary: "275,451 patients" (highlighted with a red box). To the right of the header are buttons for "New Query", "Databases", and a user profile for "sharon.nirenberg". Below the header is a green "Save Query" button.

The main content area is divided into a left sidebar and a central workspace. The sidebar contains navigation options: "Find Patients", "Map", "Visualize", "Timelines", and "Patient List". The central workspace shows a tree view of concepts under "All Concepts". The "Imaging Research Warehouse" is expanded, showing "COMPUTED TOMOGRAPHY (CT)" (highlighted with a red box) and "HEAD" (highlighted with a red box). Under "HEAD", "BC CT HEAD W CONTRAST" is selected (highlighted with a red box).

The central workspace also features a "Limit to" section with a dropdown menu set to "Patients Who" and "At Least 1x". Below this, a table of filters is visible, with "COMPUTED TOMOGRAPHY (CT) - HEAD" selected (highlighted with a red box). The table includes columns for "And" and "Anytime" with "At Least 1x" selected. The "In the Same Encounter" option is also visible.

# DICOM Tag for Imaging Modality Mapped to User-Friendly Buckets

Imaging Modality	Patient Count
DIGITAL RADIOGRAPHY	1,290,161
COMPUTED TOMOGRAPHY (CT)	720,537
ULTRASOUND	588,498
MAGNETIC RESONANCE	443,414
MAMMOGRAPHY	164,279
RADIO FLUOROSCOPY	143,145
NUCLEAR MEDICINE	62,460
BONE DENSITY SCAN/ DIGITAL RADIOGRAPHY	57,089
POSITRON EMISSION TOMOGRAPHY (PET)	36,466
OTHER (IMAGING MODALITY)	29,463
TOTAL	3,535,512

# Metadata About the Digital Pathology Images on Data Ark in MSDW

- ▶ Nearly every organ system represented including lung, heart, pancreas, kidney, liver, genitourinary, gastrointestinal, hematologic, neuropathologic, etc.
- ▶ Slides represent a wide array of pathologic processes including neoplastic, developmental, Inflammatory, toxic, metabolic, genetic, degenerative, traumatic and infectious
- ▶ Staining techniques include hematoxylin and eosin (H&E), specialized stains (ex. silver, trichome) and immunohistochemistry

	Count	Anticipated Annual Growth
<b>De-identified Digital Pathology Whole Slide Images (#)</b>	~1.5 million	~1.5 million
<b>Distinct Patients (#)</b>	~191,000	
<b>Female (%)</b>	63%	
<b>Hispanic (%)</b>	19%	
<b>Size of Digital Pathology Images Data Set</b>	~1.3 PB	~1-1.5 PB

# Location of the Digital Pathology Images on Data Ark Available via Leaf

The screenshot displays the Leaf application interface. At the top, the Leaf logo is on the left, and the text "Unsaved Query 9,521 patients" is in the center. On the right, there are buttons for "+ New Query" and "Databases".

On the left sidebar, there are navigation options: "Find Patients", "Visualize", "Timelines", and "Patient List" (highlighted with a red box and the number 2).

At the top of the main content area, there are two dropdown menus: "Current Datasets (click to edit columns)" with "Basic Demographics" and "Pathology Slides" selected (highlighted with a red box and the number 3), and an "Export Data" button (highlighted with a red box and the number 5).

The main content area displays a table of patient data. The table has columns for "Person Id", "Patient Of", "Address Postal Code", "Address State", "Age", "Ethnicity", "Gender", "Language", "Marital Status", "Race", and "Religion". The first row of data is highlighted in light blue. Below the table, there are three "View details" links for each row.

On the right side of the table, there is a "Pathology Slides File Paths" column (highlighted with a red box and the number 4) containing a list of file paths for digital pathology slides.

Person Id	Patient Of	Address Postal Code	Address State	Age	Ethnicity	Gender	Language	Marital Status	Race	Religion	Pathology Slides File Paths
00091214E29F255823102417F79673538F900B409A47A22F94DD3648618F5675	Mount Sinai Health System	000	NY	52	Hispanic or Latino	FEMALE	Unknown	Unknown	No matching concept	Unknown	/sc/arion/projects/data-ark/digital_pathology_slides/aca/aca66c82-f5a0-455a-b52f-1200826a77c6.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/110/110d5164-7b10-48b8-a31e-55409803ad1.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/2cd/2cd88102-3031-432a-87fa-594aa91f73a9.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/8fc/8fc56e4b-6492-4e2c-a7e4-52e1b21b425f.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/725/7250c20a-0b70-42da-8b3f-282f62a8a01f.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/9f8/9f8d1ba4-e21a-4cc1-a31e-8af5c1bf1ca9.tiff, /sc/arion/projects/data-ark/digital_pathology_slides/895/895ecd3c-035e-4eb6-97f8-4d1fd5733033.tiff
000AB410EAC6ABE52A40EB866855685EA72892D56724903682C830CC12840AAF	Mount Sinai Health System	100	NY	42	Not Hispanic or Latino	MALE	Unknown	Unknown	White	Unknown	
000B316F470113DD78588AFD443167E562BFA64DF02CAA7E8D59B2FDF5E9D8B0	Mount Sinai Health System	067	CT	70	Hispanic or Latino	MALE	Unknown	Unknown	White	Unknown	

# Epic for Research

# Use of Epic for Clinical Trial Recruitment is Growing

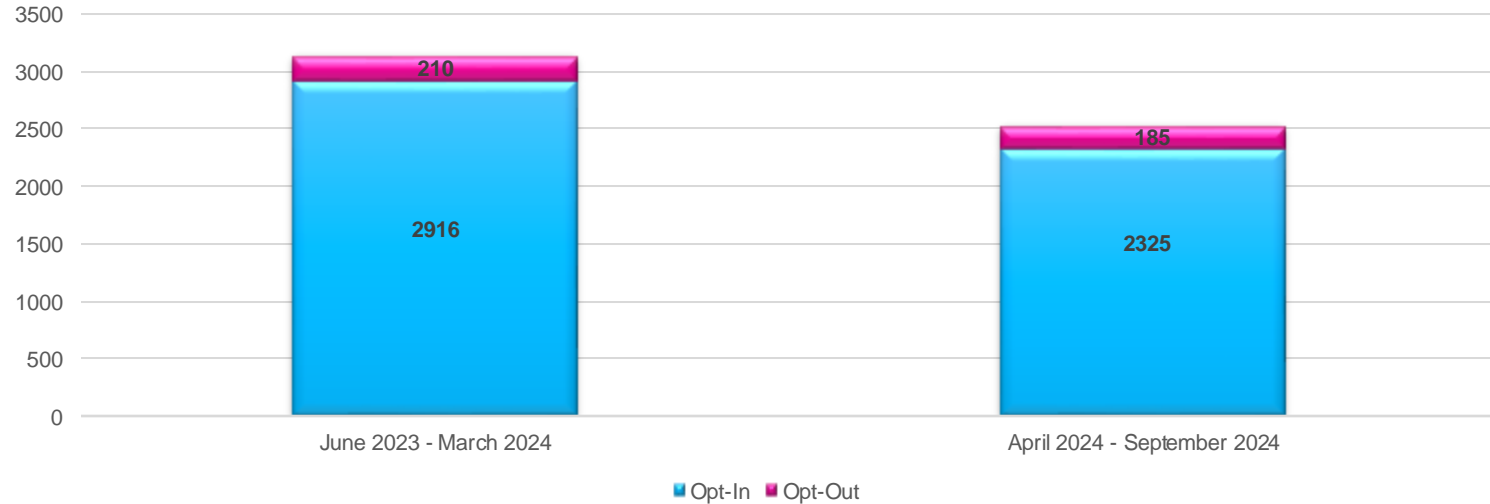
- ▶ **MyChart Recruitment and Clinical Trial Alerts are two Epic features for expanding clinical trial recruitment**
- ▶ **MyChart Recruitment**
  - Individuals are identified as potentially eligible for a clinical trial via data stored in the electronic health record
  - Patient is sent a MyChart message to alert them of potential eligibility in a study
  - The patient can express or decline interest in the study via MyChart
- ▶ **Clinical Trial Alerts**
  - Individuals are identified as potentially eligible for a clinical trial via data stored in the electronic health record
  - The provider receives an alert that a specific patient may be eligible for a study
  - Via Epic, the provider can alert the study team if the patient is interested in participating in the study

	MyChart Recruitment	Clinical Trial Alerts
Live	3	2
In progress	7	4



# Launched MyChart Research Opt-Out in June 2023

# of Patients who Opted-In and Opted-out of Being Contacted for Trial Recruitment from June 2023 to March 2024 and from April 2024 to September 2024



- ▶ In total 5,056 patients have responded to the Epic Research Consent, with 2,325 new responses since March 2024
- ▶ 634 (7.2%) of respondents have opted-out of being contacted via MyChart for research studies suggested by information in their electronic health record

# **MSDW Roadmap November 2024 –May 2025**

# MSDW Projects in Progress

	Project	Target Date	New Capabilities for Researchers
1.	<b>Upgrade MSDW OMOP ETL to the Epic Upgrade</b>	2024-Q4	Uninterrupted access to de-identified and identified MSDW OMOP database
2.	<b>Identify patients with select somatic genetic mutations in Leaf</b> <ul style="list-style-type: none"><li>Enhance clinical trial feasibility for oncology</li></ul>	2025-Q1	Enable researchers to obtain approximate counts of patients with specific somatic genetic mutations via Leaf
3.	<b>Digital Pathology</b> <ul style="list-style-type: none"><li>Pathologic diagnosis</li></ul>	2025-Q2	Enable self-service cohort identification combining pathology metadata and EHR data Enable access to 10 million de-identified pathology images on Minerva