Exploring Data Ark Data Commons: A Focus on Accessing De-identified Digital Pathology Slides Data

Yiyuan Liu, PhD
The Minerva HPC Team

March 27, 2025

Icahn
School of
Medicine at
Mount
Sinai

Outline

- Data Ark Introduction
  - Mission
  - Available Datasets
  - Data Access
  - Data Onboarding Procedures and Policy
- Accessing De-identified Digital Pathology Slides Data
  - Data Information
  - Data Access Workflow

Introduction to Data Ark

Data Ark Data Commons Increase the Power, Pace and Relevance of Our Science

Challenges

- Exhaustive searches for relevant datasets
- Repeated downloads of the same files across groups
- Difficulty in understanding opaque data structures

How Data Ark Helps

- Storage space for frequent-use research datasets
- A team managing the resources, simplifying access, training and user support

Data Ark website: https://labs.icahn.mssm.edu/minervalab/resources/data-ark/

Data Ark Offers Mount Sinai Researchers Readily Available Datasets

- There are 19 datasets hosted under Data Ark currently

Access within 24 hours after DUA signed

Public Data Sets

- 1,000 Genomes Project
- BLAST
- gnomAD
- eQTLGen
- Genebass
- GTEx
- GWAS Summary Stats
- LDSCORE
- Reference Genome
- The Cancer Genome Atlas (TCGA)
- UKBB-LD

Mount Sinai Generated Data

- CBIPM-BioMe Data
- De-identified Digital Pathology Slides
- Living Brain Project
- Mount Sinai COVID-19 Biobank
- MSDW COVID-19 EHR Data Set
- MSDW OMOP EHR Data
- STOP COVID NYC Cohort

Restricted access

User Group-acquired Data Sets

- MarketScan® (service change to be announced in April)

How to Rapidly Access Public and Mount Sinai-Generated Datasets Through Data Ark

- Visit the dataset webpage on Data Ark
- Access instruction is provided in the ‘Access’ section
- User completion of dataset-specific **DUA** (data use agreement)
- Access is granted within 24 hours

How to Access the Restricted Dataset (MarketScan) Through Data Ark

- **For MarketScan data**
  - Providing training certificates
  - Sign the DUA
  - PI signs the DUA
  - License cost applicable after initial 90-day free-of-charge access
  - Contact Dr. Inga Peter (the custodian for the MarketScan data and the MarketScan user group leader) for the cost of data access after the initial 90 days

Service change to be announced in April

Access

The following documents are required for each member of the research team:

- Mount Sinai HIPAA training certificate. Certificates of completion can be found at Mount Sinai PEAK
- CITI Basic Course and Refresher (if due) training certificates. Information for required certificates is here
- Terms of Use for MarketScan®

To use these data, you must read, agree and sign the Data Use Agreement (you must be logged in with **your Minerva ID and password** through the Mount Sinai campus network or secure remote VPN). If you don't have a Minerva ID, please open a ticket with us on MarketScan data access at hpchelp@hpc.mssm.edu

Data Onboarding Process

User-requested datasets must follow an approval process.

1. Data onboarding requested via the REDCap form (https://redcap.link/data_intake) that asks for the storage needed and prospective users.
2. Data Ark team verifies that the prospective users will use the data set.
3. If the dataset is < 1 TB, the Data Ark team will approve and start the onboarding process (webpage, copying data, verifying consent, build data usage agreement, notify users, etc.)
4. Dataset > 1 TB requires the Advisory Board for approval.

Eligibility for Cost-waived for Data Ark Hosting and Data Retention Policy

- The eligibility for cost-waived hosting on Data Ark is based on the number of user groups calibrated to the data size.

Data Size (in Terabytes)	# Of user groups/dataset	Cost waived/year
1 or less	>=2	$100
3	>=3	$300
10	>=10	$1,000 ($500 for 6 months)
20	>=20	$2,000 ($1,000 for 6 months)
30	>=20	$3,000 ($1,500 for 6 months)
100	>=20	$10,000 ($5,000 for 6 months)

- Data with annual low usage will be archived and offboarded.

Questions and Support: Contact Data Ark

All service requests must come through the ticket system:

hpchelp@hpc.mssm.edu

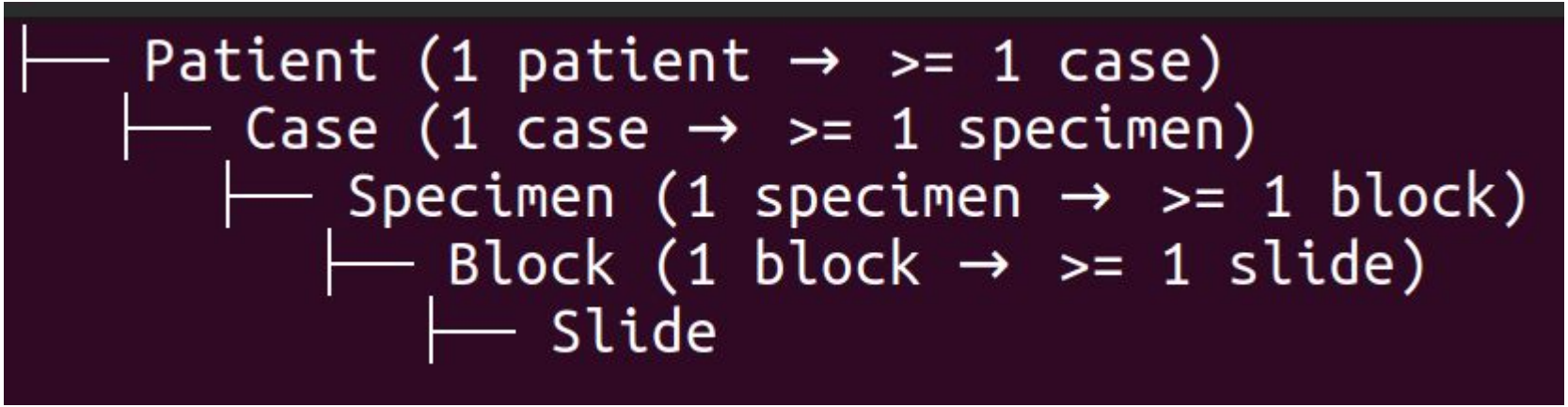
De-identified Digital Pathology Slides Data Access

De-identified Digital Pathology Slides Data:
Introduction

- In June 2019, Mount Sinai Health System (MSHS) Department of Pathology and the Philips IntelliSite Pathology Solution jointly implemented a digital pathology system, gradually replacing traditional glass slide workflows.
- By early 2024, MSHS Department of Pathology is fully digitalized.
- These specimens encompass a broad spectrum of biopsies, resections, and autopsies, reflecting the diversity of diseases affecting patients from a wide range of backgrounds.
- Almost every organ system is represented within the collection.
- The slides were prepared using a variety of staining techniques.

De-identified Digital Pathology Slides Data:
Hierarchical Pathology Imaging Assets

Multi-level Hierarchy



De-identified Digital Pathology Slides Data:
Hierarchical Pathology Imaging Assets (Continued)

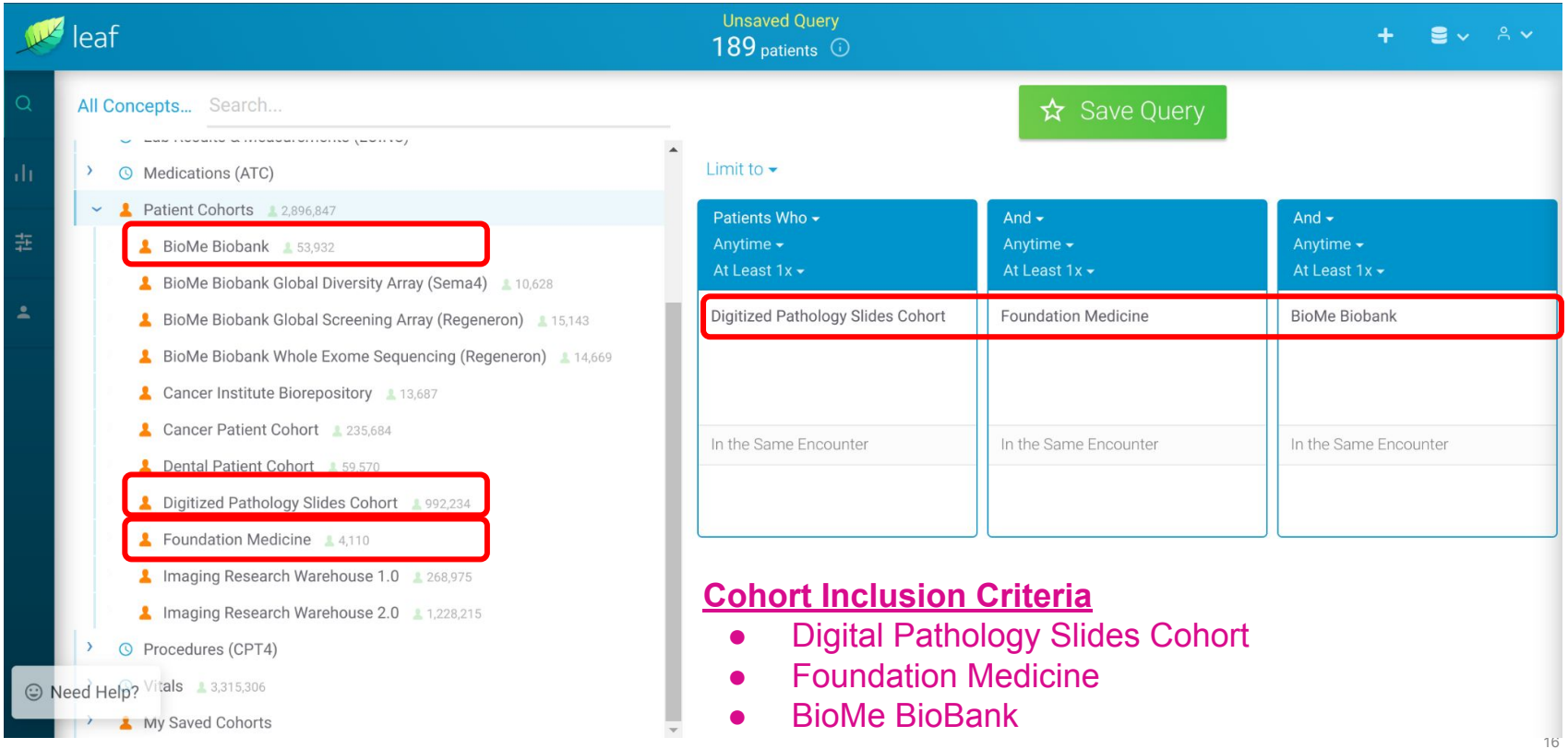
Level	Conceptual Examples
Patient	● **Patient ID**: PT-2023-001 ○ Name: Jane Smith ○ MRN: 123456789 ○ Demographics, clinical history
Case	● **Case ID**: C-2023-001 ○ Diagnosis: Invasive ductal carcinoma (Breast) ○ Date: 2023-05-08 ○ Pathologist: Dr. Lee
Specimen	● **Specimen ID**: SP-2023-001-A ○ Type: Core needle biopsy ○ Laterality: Left breast ○ Collection date: 2023-05-10
Block	● **Block ID**: BL-2023-001-A1 ○ Derived from: SP-2023-001-A ○ Processing date: 2023-05-12
Slide	● **Slide ID**: SL-2023-001-A1-1 ○ Stain: H&E ○ Scan date: 2023-05-13

De-identified Digital Pathology Slides Data:
Electronic Health Records (EHR) and Slide Image Data

- EHR data is served through Mount Sinai Data Warehouse.
- Slide image data
  - De-identified and accessible via Data Ark on Minerva
  - TIFF format
  - 40x resolution (~0.25 microns/pixel)
  - > 1.6 million slides with new slides ingested daily
  - Retrospective slide digitization has not yet been performed.
- With an approved IRB protocol, crosslinking of EHR and slide image data is provided as a service by Mount Sinai Data Warehouse.

Build a Cohort in Leaf:

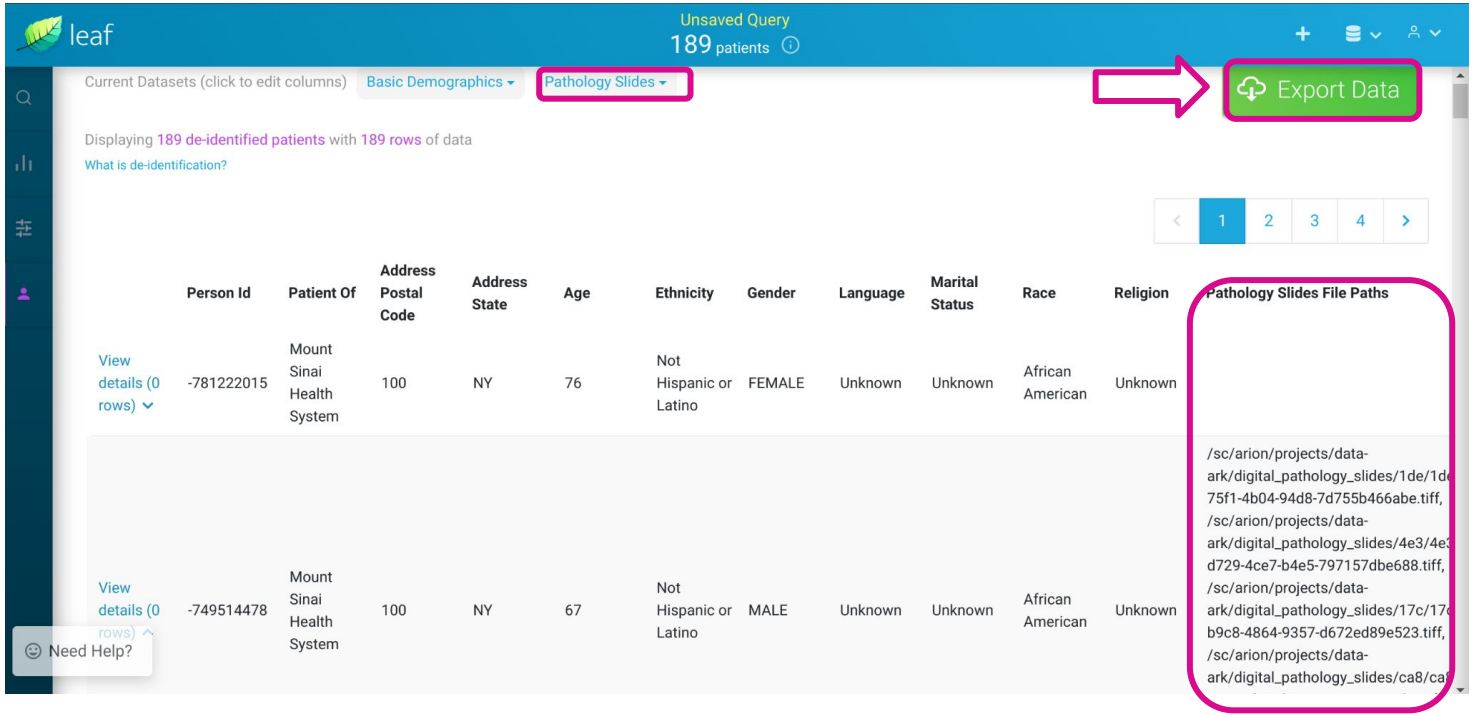
https://leaf.mssm.edu



Cohort Inclusion Criteria

- Digital Pathology Slides Cohort
- Foundation Medicine
- BioMe BioBank

Pathology Slides Minerva Paths Are Accessible on Leaf



Access Pathology Slides on Minerva

- Complete data use agreement (DUA: **https://dataarkforms.hpc.mssm.edu/**).
- Access grant confirmations are delivered via email within 24 hours.
  - Over 1.6 million slides are organized in subfolders of the data collection **/sc/arion/projects/data-ark/digital_pathology_slides**.
  - Alternatively, load the Data Ark module ("**ml dataark**") and the slide collection is referenced in the environment variable **$DPslides**.
- Data Ark team is **in the process of deploying Digital Slide Archive** (**https://digitalslidearchive.github.io/**), a slide viewer for visualizing and annotating over 1.6 million slides hosted at Data Ark.

Contact for Help

- **Inquiries of Digital Pathology Slides access or other Data Ark/Minerva service**
  - HPC ticket system **hpchelp@hpc.mssm.edu**.
- **EHR data access, Leaf or related inquiries**
  - Mount Sinai Data Warehouse ticket system **https://labs.icahn.mssm.edu/msdw/open-a-ticket/**.
- **Computational pathology modeling related collaborations**
  - Dr. Gabriele Campanella (PI), Windreich Department of AI and Human and Health email: **gabriele.campanella@mssm.edu**.

Please Acknowledge CTSA in Your Publications

- Please acknowledge the support from Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai by including the following acknowledgement in a publication of any material, whether copyrighted or not, based on or developed with Minerva HPC resources:

  "This work was supported in part through the computational resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences."

CTSA Clinical & Translational Science Awards ®


DATA ARK


Thank You!