

AIR•MS Training Session III: Multimodal Data

Andrew Deonarine

Ashwin Sawant

Credits: Eugenia Allewa

Sept 24st, 2026



Icahn School
of Medicine at
**Mount
Sinai**

Agenda

- 1) Recap – OMOP and AIR·MS
- 2) Overview of AIR·MS data
- 3) Pathology
- 4) Radiology
- 5) Mount Sinai Million
- 6) Cardiovascular Imaging
- 7) Summary



Some Additional Resources

Getting Started Workbooks:

<https://github.com/scdmsanalytics-bit/airms-training>

scdmsanalytics-bit Delete AIRMS-Training.ipynb 182f32f · last month 4 Commits		
AIRMS-Training-Session-1.ipynb	Updated with dotenv	last month
AIRMS-Training-Session-2.ipynb	Updated with dotenv	last month
LICENSE	Initial commit	last month
README.md	Initial commit	last month

Tutorials for Various Schemas (Python, R):

<https://github.mountsinai.org/AIRMS/airms-researcher-tutorials-minerva>

main airms-researcher-tutorials-minerva / AIR-MS_Tutorials / Go to file	
Anurag Patil and Anurag Patil Added Safety dependency check cell 6bfbb23 2 days ago History	
..	
CDMECG	Added Safety dependency check cell 2 days ago
CDMECHO	Added Safety dependency check cell 2 days ago
CDMGIDATABASE	Added Safety dependency check cell 2 days ago
CDMICU	Added Safety dependency check cell 2 days ago
CDMMSM	Added Safety dependency check cell 2 days ago
CDMPATHOLOGY	Added Safety dependency check cell 2 days ago
CDMRADIOLOGY	Added Safety dependency check cell 2 days ago
Data_Dictionary	Added Safety dependency check cell 2 days ago
MSDW_OMOP_DE-ID	Added Safety dependency check cell 2 days ago
MSDW_OMOP_PHI	Added Safety dependency check cell 2 days ago
OMOP_SYNPUF_5%	Added Safety dependency check cell 2 days ago

Recap

What is AIR·MS?

Artificial Intelligence-Ready Mount Sinai (AIR·MS) is a research platform that is composed of:

- 1) A (very fast) integrated database including Mount Sinai Data Warehouse (MSDW), Pathology and Radiology metadata; and the included data is growing.
- 2) A Research Environment that allows interactions with the AIR·MS database from Python or R.
- 3) An Application Tier to host a growing number of applications – including cohort building tools and annotation apps.

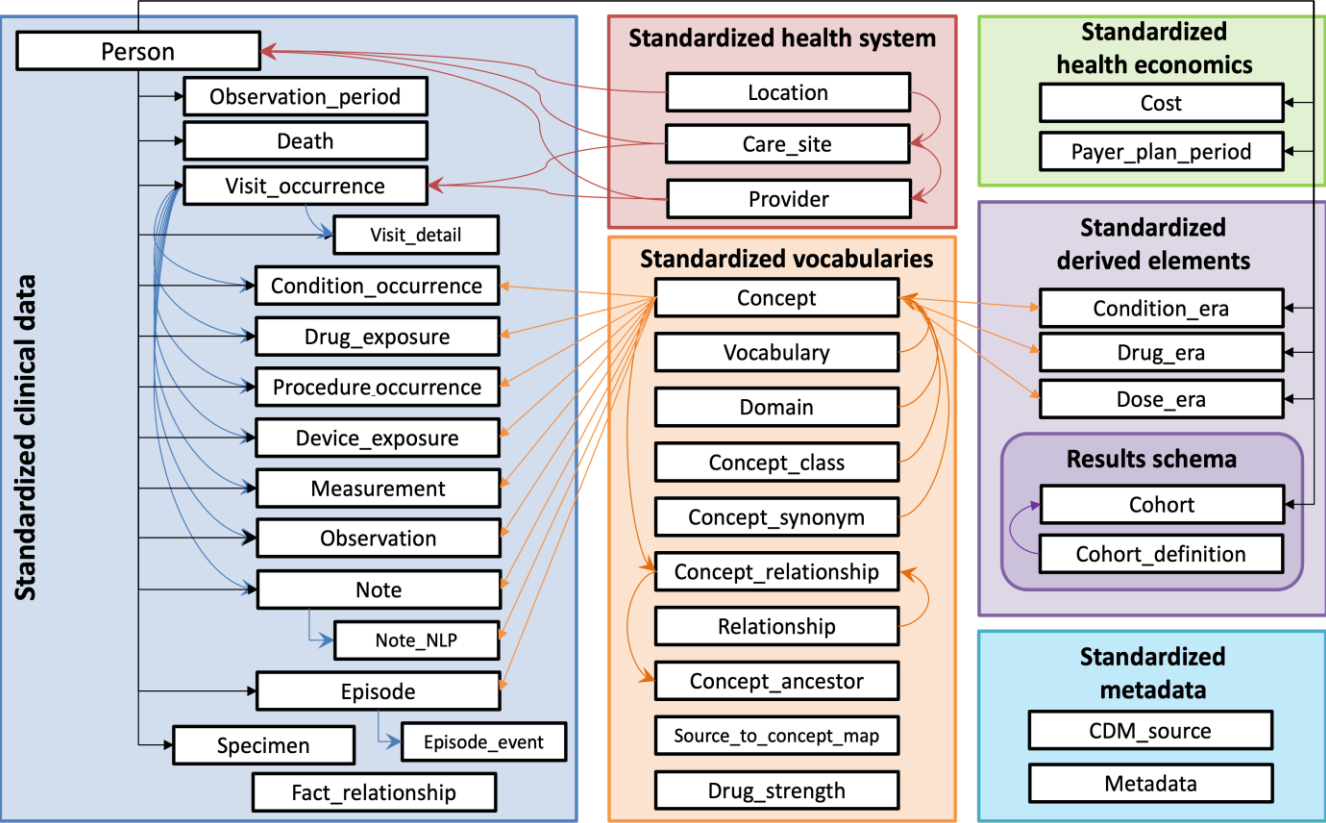


<https://labs.icahn.mssm.edu/minervalab/air-ms-artificial-intelligence-ready-mount-sinai/>

Epic, MSDW, and AIR-MS



Main Tables in OMOP Database



Dataset	OMOP Table	# Records
Patients	person	12,051,324 patients
Encounters	visit_occurrence	111,804,681 encounters
Diagnosis	condition_occurrence	120,323,291 diagnoses
Lab Results	measurements	1,107,913,246 results
Medication	drug_exposure	138,120,557 prescriptions
Procedures	procedure_occurrence	326,372,912 procedures
Observations	observation	145,273,185 social histories
Notes	note	204,925,569 notes
Providers	provider	601,032 specialists
Care Sites	care_site	82,103 beds

As of April, 2025

Mapping Concepts and Information

From session 1: we have mapping between the coding systems in Epic and the coding systems used by OMOP:

- We can't just use one giant, static mapping table of Epic codes to OMOP codes because some of the coding systems have commercial licenses.
- Don't just perform keyword searches on codes: sometimes the keywords don't appear, consult an expert clinician.
- (ICD = International Classification of Diseases, SNOMED = Systematized Nomenclature of Medicine, NDC = National Drug Code, ATC = Anatomical Therapeutic Classification, LOINC = Logical Observation Identifiers Names and Codes, CPT = Current Procedural Terminology)

Blue =
License /
Proprietary
Coding



Diagnosis

Epic



Epic / ICD-10  SNOMED-CT
Custom



Drug

NDC, ATC  RxNorm
Athena



Labs

LOINC  LOINC



Procedures

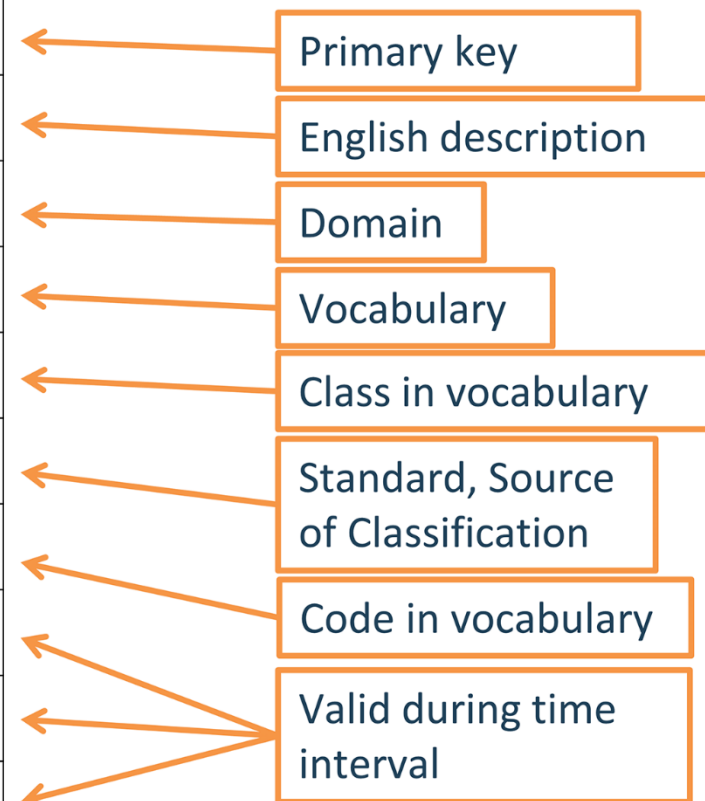
CPT

 SNOMED-CT
Custom

Using the Data: Querying and Retrieving Data

- Foundational to OMOP is a “Concept” (stored in the CONCEPT table)
- Concept domains: “Condition,” “Drug,” “Procedure,” “Visit,” “Device,” “Specimen,” etc.

CONCEPT_ID	313217
CONCEPT_NAME	Atrial fibrillation
DOMAIN_ID	Condition
VOCABULARY_ID	SNOMED
CONCEPT_CLASS_ID	Clinical Finding
STANDARD_CONCEPT	S
CONCEPT_CODE	49436004
VALID_START_DATE	01-Jan-1970
VALID_END_DATE	31-Dec-2099
INVALID_REASON	



OMOP Tables

- Have a common naming structure for columns, tables, and can link tables using various IDs

person Column	Description
person_id	Person identifier
gender_concept_id	Gender
race_concept_id	Race
ethnicity_concept_id	Ethnicity
location_id	Location
provider_id	Provider
gender_source_value	(from source)
race_source_value	(from source)
...	

visit_occurrence Column	Description
visit_occurrence_id	ID
person_id	Person ID
visit_start_date	Date
care_site_id	Location
provider_id	Provider
...	

condition_occurrence Column	Description
condition_occurrence_id	ID
person_id	Person ID
condition_concept_id	Concept ID
condition_start_date	Start
condition_end_date	End
stop_reason	Why condition no longer present...
...	

<https://www.ohdsi.org/>

Patients (person Table)

observation		Person	Location		Death	
Patient A	Race	Patient A	Patient A	Home Address A	Patient A	Death date
	Ethnicity					
	Language Preference					
	Sexual Orientation					
Patient B	Race	Patient B	Patient B	Home Address B		
	Marital Status					
	Gender Identity					
Patient C	Ethnicity	Patient C				
	Religious Affiliation					

person Column	Description
person_id	Person identifier
gender_concept_id	Gender
race_concept_id	Race
ethnicity_concept_id	Ethnicity
location_id	Location
provider_id	Provider
gender_source_value	(from source)
race_source_value	(from source)
...	

<https://www.ohdsi.org/>

Overview of AIR-MS Data

Different Datasets Stored in AIR·MS

Currently Available Modalities:

- Mount Sinai Data Warehouse (MSDW), both containing protected health information (PHI) and DeID (deidentified) Observational Medical Outcomes Partnership (OMOP)-mapped electronic health record (EHR)
- Pathology Metadata
- Radiology Metadata
- BioMe/Sinai Million
- Electrocardiogram (EKG)
- Echocardiography

Work in progress: GI Research Database, electroencephalogram (EEG), Endoscopy & Colonoscopy Reports

All modalities are stored in separate database schemas, and access is granted to each schema individually based on Institutional Review Board (IRB)

Pathology

Slide Credits: Gabriele Campanella, PhD

Assistant Professor, Artificial Intelligence and Human Health

Pathology Lab Workflow and Data

This work was supported by

- Department of Pathology, Molecular and Cell-Based Medicine
- Windreich Department of Artificial Intelligence and Human Health
- Scientific Computing and Data
- Digital and Technology Partners



De-identified Digital Pathology Training Wednesday, April 22 | 10–11 am



Mount Sinai is Providing the Largest Set of Slides Available for Research in the Country



Each slide is linked to multiple sets of clinical data



Featured organ systems include lung, heart, pancreas, kidney, liver, gastrointestinal, genitourinary, gynecological, hematological, and neuropathological systems

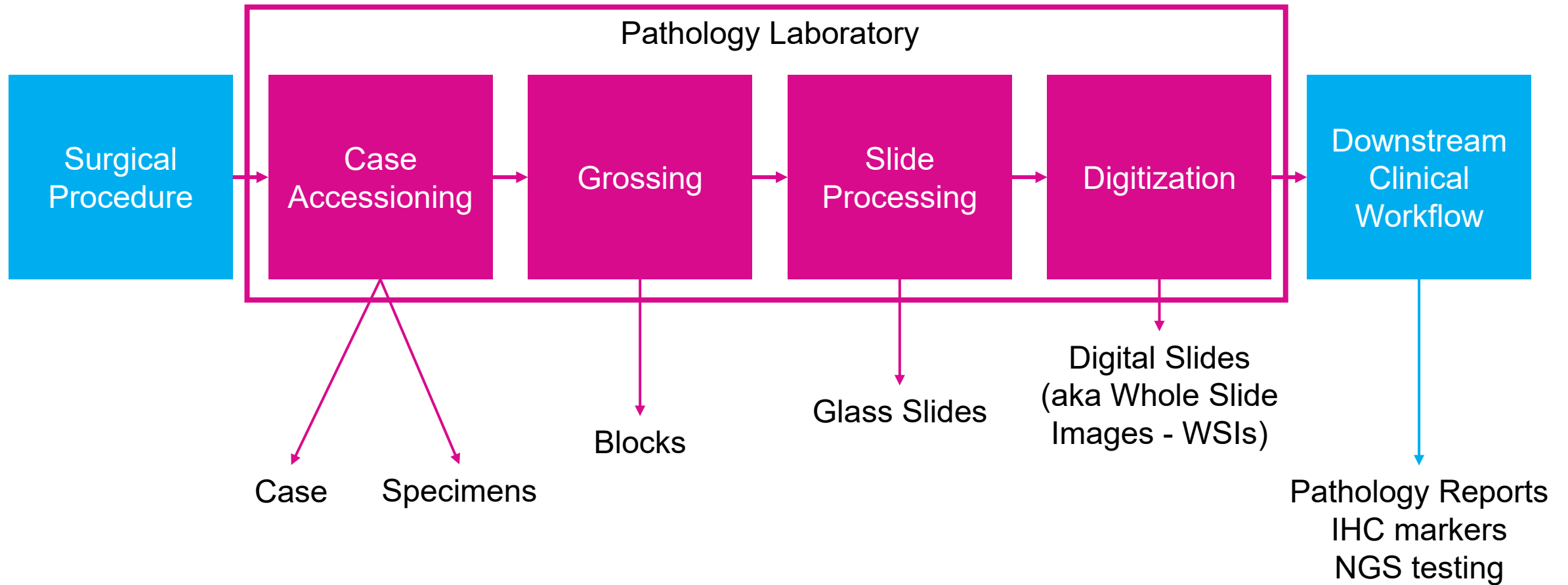


Featured disease processes include neoplastic, inflammatory, genetic, degenerative, traumatic, and infectious pathologies

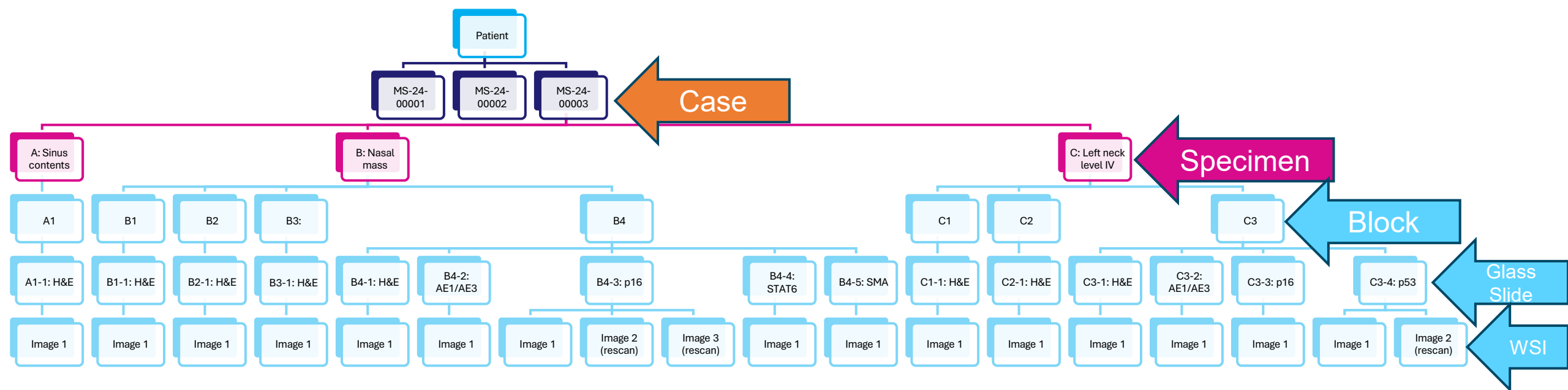


**Scan the QR code
to join this training session**

Pathology Lab Workflow and Data



Pathology Data Hierarchy



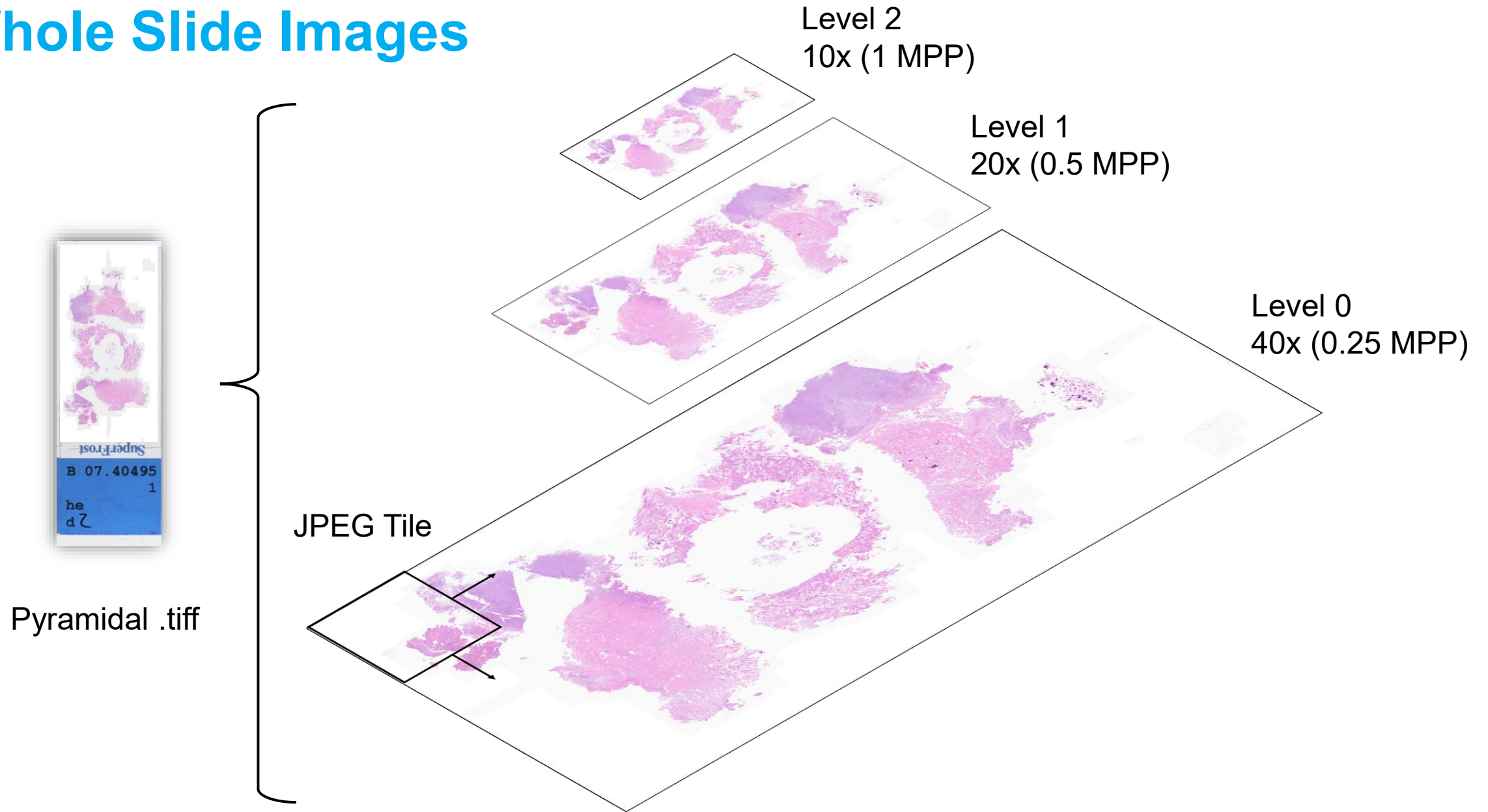
Overview of Pathology Data

The pathology department produces large amounts of physical and digital data.

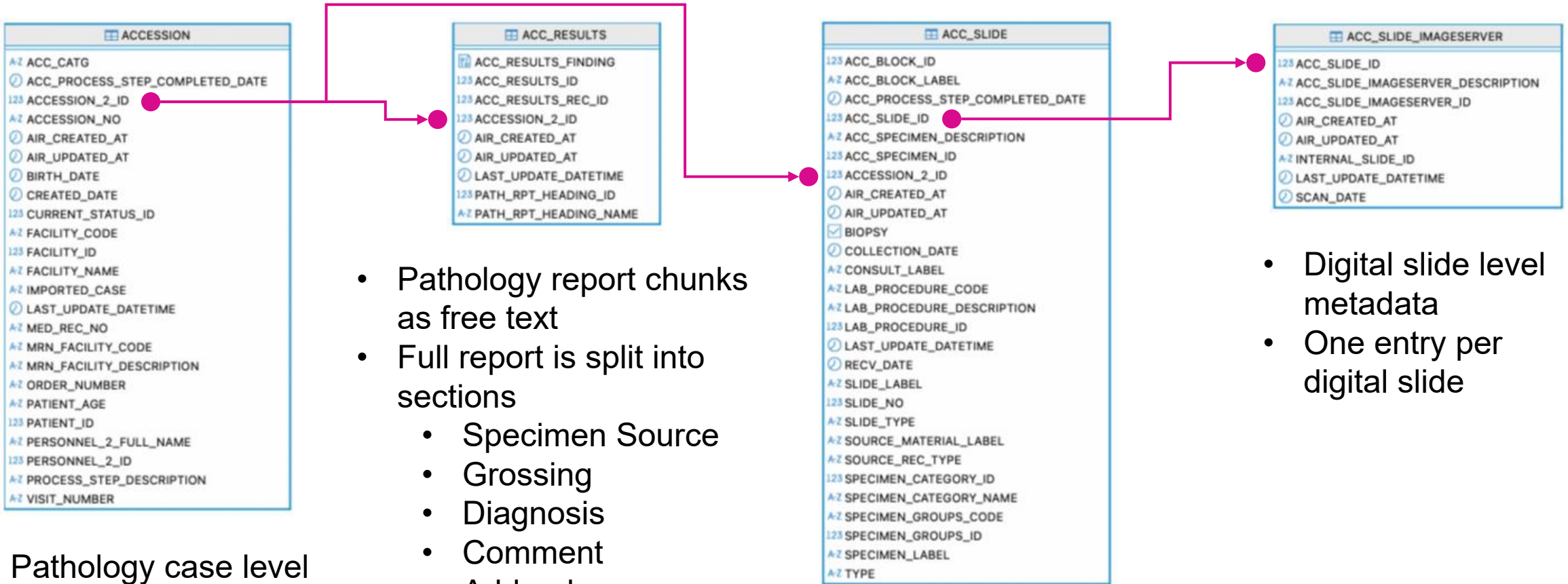
Data Modalities

- Biological Assets → Stored for a minimum of 20 years per state regulations
 - Tissue blocks
 - Glass Slides
- Digital Assets
 - Pathology Laboratory Information System (LIS) Metadata → Links cases, specimens, blocks, slides to reports. Tracks most aspects of pathology lab
 - Whole Slide Images → Pyramidal .tiff files scanned with Philips WSI scanners at 40x resolution. Available as part of Data Ark
 - Pathology Reports → Free text diagnosis written by the pathologist for the entire case. Can include test results (IHC, FISH, etc.) as free text.

Whole Slide Images



Pathology Data in AIR-MS



- Pathology case level metadata
- One entry per case

- Pathology report chunks as free text
- Full report is split into sections
 - Specimen Source
 - Grossing
 - Diagnosis
 - Comment
 - Addendum
- One entry per section

- Glass slide level metadata
- One entry per glass slide

- Digital slide level metadata
- One entry per digital slide

AIRMS Schema:
CDMPATHOLOGY

Case Metadata

Frequently used fields:

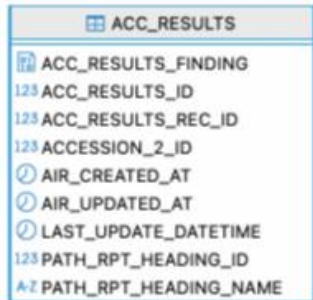
- *ACCESSION_2_ID* Linking key to other tables.
- *ACCESSION_NO* Accession number, used to identify cases. Pathologist-friendly.
- *MED_REC_NO* EPIC MRN. Can be used to link to other data modalities.

ACCESSION	
A-Z	ACC_CATG
②	ACC_PROCESS_STEP_COMPLETED_DATE
123	ACCESSION_2_ID
A-Z	ACCESSION_NO
②	AIR_CREATED_AT
②	AIR_UPDATED_AT
②	BIRTH_DATE
②	CREATED_DATE
123	CURRENT_STATUS_ID
A-Z	FACILITY_CODE
123	FACILITY_ID
A-Z	FACILITY_NAME
A-Z	IMPORTED_CASE
②	LAST_UPDATE_DATETIME
A-Z	MED_REC_NO
A-Z	MRN_FACILITY_CODE
A-Z	MRN_FACILITY_DESCRIPTION
A-Z	ORDER_NUMBER
A-Z	PATIENT_AGE
123	PATIENT_ID
A-Z	PERSONNEL_2_FULL_NAME
123	PERSONNEL_2_ID
A-Z	PROCESS_STEP_DESCRIPTION
A-Z	VISIT_NUMBER

Report Text

Frequently used fields:

- *ACCESSION_2_ID* Linking key to case table.
- *PATH_RPT_HEADING_NAME* Name of the section. Section naming is not standardized. Some examples:
 - DIAGNOSIS
 - FINAL DIAGNOSIS
 - SPECIMEN SOURCE
 - NOTE
 - ADDENDUM 1, 2, 3, ...
 - COMMENT
 - GROSSING
- *ACC_RESULT_FINDING* Actual report text.



ACC_RESULTS	
ACC_RESULTS_FINDING	
ACC_RESULTS_ID	
ACC_RESULTS_REC_ID	
ACCESSION_2_ID	
AIR_CREATED_AT	
AIR_UPDATED_AT	
LAST_UPDATE_DATETIME	
PATH_RPT_HEADING_ID	
PATH_RPT_HEADING_NAME	

Glass Slide Metadata

ACC_SLIDE	
123	ACC_BLOCK_ID
A-Z	ACC_BLOCK_LABEL
②	ACC_PROCESS_STEP_COMPLETED_DATE
123	ACC_SLIDE_ID
A-Z	ACC_SPECIMEN_DESCRIPTION
123	ACC_SPECIMEN_ID
123	ACCESSION_2_ID
②	AIR_CREATED_AT
②	AIR_UPDATED_AT
☒	BIOPSY
②	COLLECTION_DATE
A-Z	CONSULT_LABEL
A-Z	LAB_PROCEDURE_CODE
A-Z	LAB_PROCEDURE_DESCRIPTION
123	LAB_PROCEDURE_ID
②	LAST_UPDATE_DATETIME
②	RECV_DATE
A-Z	SLIDE_LABEL
123	SLIDE_NO
A-Z	SLIDE_TYPE
A-Z	SOURCE_MATERIAL_LABEL
A-Z	SOURCE_REC_TYPE
123	SPECIMEN_CATEGORY_ID
A-Z	SPECIMEN_CATEGORY_NAME
A-Z	SPECIMEN_GROUPS_CODE
123	SPECIMEN_GROUPS_ID
A-Z	SPECIMEN_LABEL
A-Z	TYPE

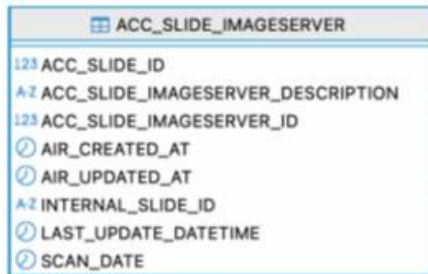
Frequently used fields:

- *ACCESSION_2_ID* Linking key to case table.
- *ACC_SLIDE_ID* Linking key to other tables.
- *SPECIMEN_LABEL* Specimen identifier within the case. Uppercase letter (A, B, ...). Matches specimen labels in free text pathology report. Pathologist-friendly.
- *ACC_BLOCK_LABEL* Block identifier within the specimen. Number (1, 2, ...). Matches block labels in free text pathology report, usually in the grossing description. Pathologist-friendly.
- *SLIDE_LABEL* Slide identifier within the block. Number (1, 2, ...). Not used in free text reports usually. Pathologist-friendly.
- *RECV_DATE* Received date of the specimen. Useful to match cases to other data modalities.
- *LAB_PROCEDURE_DESCRIPTION* Lab order that generated the slide. Usually a stain name. For example: H&E, PDL1, ER, PR, ...

Digital Slide Metadata

Frequently used fields:

- *ACC_SLIDE_ID* Linking key to slide table.
- *INTERNAL_SLIDE_ID* 6-char alphanumeric code that uniquely identifies a glass slide. Not pathologist-friendly. Aka BARCODE. If a barcode exists, the slide was scanned.
- *ACC_SLIDE_IMAGESERVER_DESCRIPTION* long alphanumeric code that uniquely identifies a .tiff file. Unique to a scanning event, not to a glass slide. Same barcode can be associated to multiple .tiff files (e.g., rescanning). Not pathologist-friendly.
- *SCAN_DATE* Timestamp of scanning event.



	ACC_SLIDE_IMAGESERVER
125	ACC_SLIDE_ID
A-Z	ACC_SLIDE_IMAGESERVER_DESCRIPTION
125	ACC_SLIDE_IMAGESERVER_ID
🕒	AIR_CREATED_AT
🕒	AIR_UPDATED_AT
A-Z	INTERNAL_SLIDE_ID
🕒	LAST_UPDATE_DATETIME
🕒	SCAN_DATE

NOTE:

- Scanning started in 2020 at 40% capacity. 100% prospective digitization was reached in 2024. Most slides are still not scanned.
- *INTERNAL_SLIDE_ID* and *ACC_SLIDE_IMAGESERVER_DESCRIPTION* can be used to identify slides in Data Ark. Mapping is not publicly available.

Pathology Metadata – Practical Tips

- **How to communicate with the pathologist**

Accession numbers, specimen/block/slide labels are well understood by pathologists.

For example: slide “A1-1” from block “A1” in case MS-24-00001.

- **Linking slides to a diagnosis is challenging**

Reports are written for the entire case. Most granular diagnosis is at the specimen level (unstructured).

There is no slide level description. It is important to talk to a pathologist to understand which slides from a case/specimen are relevant to your study.

- **Some metadata can be inaccurate**

Some metadata fields exposed are used internally in the pathology department for workflow management. They are not for diagnostic purposes and can be inaccurate. Examples include: division, organ, specimen description, ICD codes. Always talk to a pathologist to discuss how to best approach your problem.

- **Data Ark slide links are not publicly available**

To obtain Minerva paths to a set of slides can be done through MSDW (requires IRB). You will need *INTERNAL_SLIDE_ID* and *ACC_SLIDE_IMAGESERVER_DESCRIPTION* for each slide in your cohort.

Radiology

Slide Credits: Francesco La Rosa, PhD
Instructor, Artificial Intelligence and Human Health

Overview of Radiology data

What type of data is it, and where is it produced?

- Medical imaging data (e.g., MRI, CT, X-ray) produced by radiology scanners and associated acquisition systems. During routine clinical care it is acquired for diagnosis and treatment planning.

Where is it stored?

- In PACS for clinical use, in XNAT for research studies.

What data standards exist?

- DICOM: how medical images and associated metadata are formatted, stored, and annotated across systems. It ensures interoperability between scanners, PACS, and analysis tools by specifying image encoding, patient/study identifiers, acquisition parameters, and standardized metadata tags.
- DICOM tags: identify the attribute, usually in the format (XXXX,XXXX) with hexadecimal numbers. The first group represents the **DICOM group number**, and the second represents the **element number**. Together, they uniquely define each tag. For example, (0008,0060) for Modality or (0010,0010) for Patient Name.
- DICOM tag explorer: <https://www.dicomlibrary.com/dicom/dicom-tags/>

Radiology Metadata on AIR·MS

Schema and Access

- Stored within the CDMRADIOLOGY schema in AIR·MS.
- Access requires: IRB approval and project registration. Can be request in SailPoint.

Tables and Content

- Two tables with several attributes available on AIR·MS

RADIOLOGY_METADATA
ID
PATIENT_ID
SERIES_INSTANCE_UID
STUDY_INSTANCE_UID
ETL_RECORD_UPDATE_DATETIME

RADIOLOGY_DICOM_DATA
ID
RADIOLOGY_METADATA_ID
AG_VALUE_REPRESENTATION
DICOM_TAGS
SERIES_INSTANCE_UID
STUDY_INSTANCE_UID
TAG_XTN_PATIENT_EPIC_MRN

Radiology Metadata on AIR·MS

Linkage to EHR Cohorts

- The RADIOLOGY_METADATA **PATIENT_ID** correspond to the MRN codes and can be linked with OMOP EHR dataset in AIRMS.

Raw Data Storage

- The imaging data is stored in PACS
- It can be requested from the Mount Sinai Imaging Warehouse
- De-identification of images in progress, aim to have 5 years of CT/MRI studies in 2026

Radiology Metadata - Pitfalls and Dangers

Common mistakes by non-experts

- One study (STUDY_INSTANCE_UID) can contain many SERIES_INSTANCE_UIDs (each corresponding to a different sequence, acquisition, or reconstruction).

Other important aspects

- The DICOM tags values are not always consistent, so they should always be verified and standardized before analysis.

Radiology Metadata Use Case

Hypothetical use case:

- Identify all adults diagnosed with spondylolisthesis who subsequently underwent fusion surgery and had at least one pre- or post-operative imaging study.

Data sources and Linkage strategy

- OMOP tables PERSON, CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, and CONCEPT_ANCESTOR to identify spondylolisthesis and fusion events
- Linked to imaging data in CDMRADIOLOGY.RADIOLOGY_METADATA and CDMRADIOLOGY.RADIOLOGY_DICOM_DATA via MRN (PATIENT_ID) and metadata IDs
- Filtered for MRI lumbar studies using DICOM Modality and Body Part Examined tags.

Mount Sinai Million/BioMe Biobank

Slide Credits: Michael Preuss, PhD

Assistant Professor, Artificial Intelligence and Human Health

Overview of Mount Sinai Million Health Discoveries Program

What type of data is it?

- Meta data, sample IDs for the Mount Sinai Million Health Cohort (BioMe Biobank, MSM Biobank)
 - MRNs (PHI data) &
 - De-id MRNs
- Same sample for actual genetic data, MSM Biobank (BioMe Biobank) on Minerva

Where is it stored?

- The genetic data is stored on Minerva HPC in /sc/arion/projects/MSM

Access via registration [Data and Specimen form](#)

What data standards exist, (only Minerva)?

- Imputation (bgen format) & WES data VCF format
- Single sample gVCF
- PhecodeX tables
- Annotated data (VEP, including LOFTEE and dbNSFP)
- Covariate file with genetic principal components, genetic ancestry, batch covariates, age and sex
- Bulk datasets with QC metrics and CRAM files

Mount Sinai Million / BioMe on AIR-MS (PHI / de-id) 1/2

Schema and Access

Stored within the schema

- CDMMSM (PHI, MRN)
- CDMMSMDEID (de-id)

Access requires:

- IRB approval (only CDMMSM) in SailPoint Mount Sinai Million / BioMe identifiable &
- (optional) project registration [Data and Specimen form](#) for actual genetic data access on Minerva

Linkage to EHR Cohorts

- Linkage is performed via the Medical Record Number (MRN):
 - `XTN\_PATIENT\_EPIC\_MRN` in the PERSON table of the MSDW OMOP schema (CDMPHI)
 - `MRN` in the PATIENT table of the MSM Biobank schema (CDMMSM)

Raw data storage

- On Minerva `/sc/arion/projects/MSM`

Mount Sinai Million / BioMe on AIR·MS (PHI / de-id) 2/2

Tables and Content

Mount Sinai Million / BioMe identifiable (PHI) *

Schema: CDMMSM Data •
snapshot from 06/09/2025 •
Unique patients: 279,885

*Schema: CDMMSMDEID (DE-ID)
same but without MRN

PATIENT	Comments
ID	Internal ID for linking to other tables within the dataset
MRN	Medical Record Number (EPIC MRN - only accessible under regulatory approval)
MASKED_MRN	De-Identifier for combined BioMe Biobank set with Regeneron and Sema4 data
RGN_ID	De-identifier for first Regeneron batch regarding BioMe Biobank
SEMA4_ID	De-identifier for Sema4, a subset of Masked MRN ID
MSM_ID	De-identifier for Mount Sinai Million Biobank, a combined setoff RGN_ID and new MSM ID
MILLION_ID	Indicator for all consented patients with and without genomic data
AIR_CREATED_AT	Record creation in AIR·MS
AIR_UPDATED_AT	Record updated in AIR·MS

Medical Record Number (PHI) serves to cross Reference with OMOP and other data modalities

Only of importance when you were working with BioMe Biobank and And want to cross reference MSM Biobank

MSM de-id for the Mount Sinai Million Biobank cohort, sample identifier (patient was genotyped)

Just an indicator ID, its presence shows that the patient was consented by the MSM protocol (the patient can be genotyped or not (yet))

Pitfalls and Dangers

Common mistakes by non-experts

- What are the typical errors you see when people new to this data modality start working with it?
- Examples: misinterpretation of variables, misuse of identifiers, incorrect assumptions about completeness.

Other important aspects

- What should users keep in mind when analyzing or linking this modality?
- Are there modality-specific caveats (e.g. unit conversions, coding systems)?

General pitfalls

- What traps do people often fall into that could compromise data quality, analysis, or interpretation?

CDMMSM Use Case

Hypothetical use case:

“Identify Mount Sinai Million participants with a hypertension diagnosis (ancestor concept 320128), returning their earliest diagnosis date along with Person ID and MRN, restricted to consented individuals present in CDMMSM.PATIENT.”

Data sources and linkage strategy:

OMOP tables: person, condition_occurrence, measurement (for blood pressure), procedure_occurrence (for treatment codes).

Biobank tables: CDMMSM.PATIENT (MSM_ID, MRN, consent indicators).

Linkage identifier: EPIC Medical Record Number (XTN_PATIENT_EPIC_MRN ↔ MRN).

Strategy: Restrict to consented MSM participants; subset to those with non-null MSM_ID (genetic data available).

Sample Query

- Extracts patients with a hypertension diagnosis (using OMOP ancestor concept 320128), returning their earliest diagnosis date (DX_DATE), Person ID, and MRN.
- Only includes patients who are part of the Mount Sinai Million (consented cohort), identified by the presence of their MRN in CDMMSM.PATIENT.

```
SELECT
    MIN(co.condition_start_date) AS dx_date,
    co.person_id,
    pe.xtn_patient_epic_mrn AS mrn
FROM CDMPHI.condition_occurrence AS co
JOIN CDMPHI.concept_ancestor AS ca
    ON ca.descendant_concept_id = co.condition_concept_id
JOIN CDMPHI.person AS pe
    ON pe.person_id = co.person_id
JOIN CDMMSM.patient AS p
    ON p.mrn = pe.xtn_patient_epic_mrn
WHERE ca.ancestor_concept_id = 320128
AND pe.xtn_patient_epic_mrn IS NOT NULL
GROUP BY
    co.person_id,
    pe.xtn_patient_epic_mrn
```

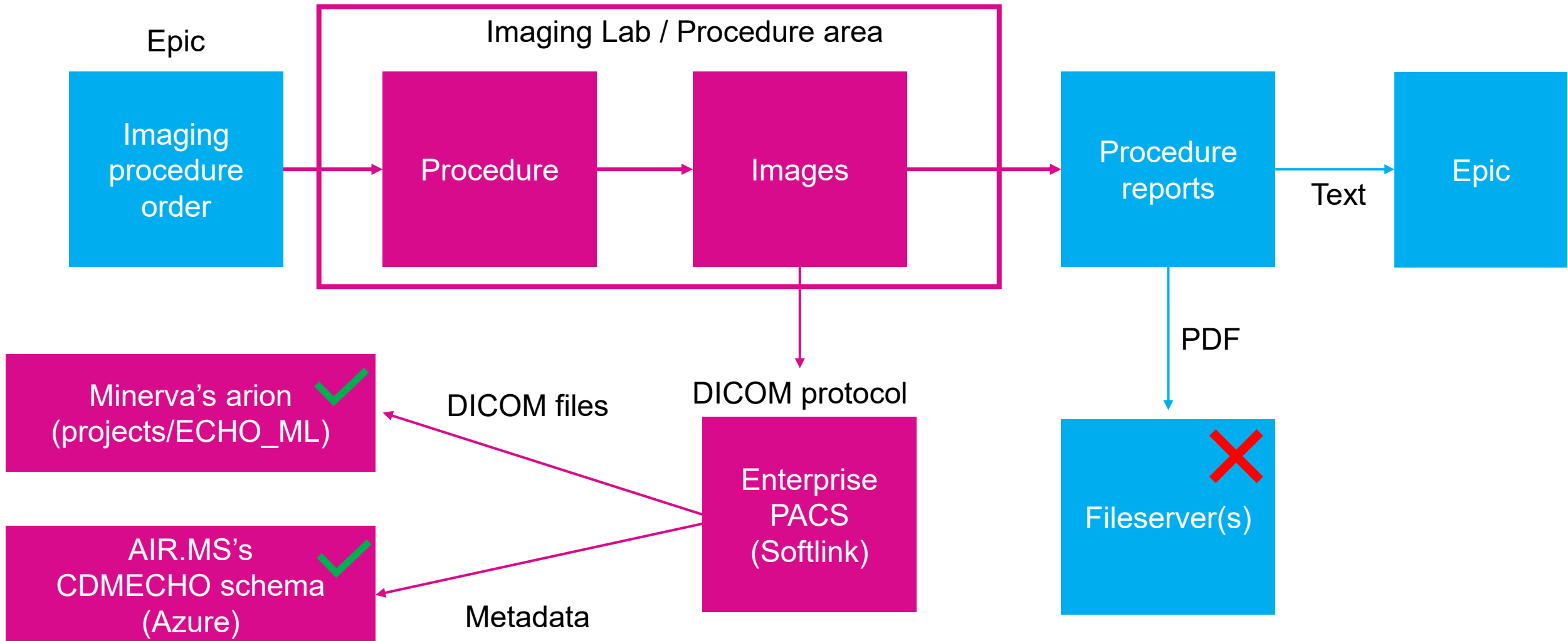
NB: can substitute 261326 with any other standard OMOP concept to change the query for another disorder (and including all child concepts). Find standard concepts [here](#)

Cardiovascular Imaging

Ashwin Sawant, MD

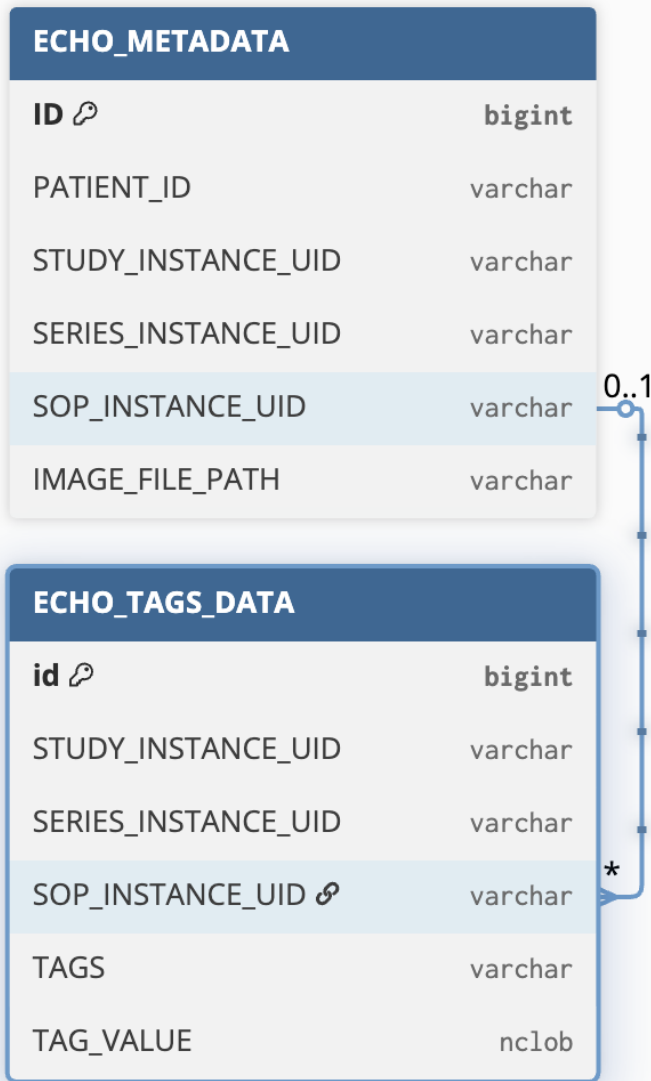
Assistant Professor, Artificial Intelligence and Human Health

Overview of cardiovascular imaging



Types of imaging studies

- Echocardiograms
- Coronary artery catheterization
- Transcatheter aortic valve replacement
- Peripheral vascular procedures
- We also found some CT, MRIs but haven't explored them



Tags of interest:

1. PatientID: MRN

2. StudyDate, StudyTime

3. Modality:

1. XA = x-ray angiography (cardiac catheterization, peripheral vascular catheterization)

2. US = ultrasound (echocardiography)

4. IMAGE_FILE_PATH:

/sc/arion/projects/ECHO_ML/softlink_archive/...

SAMPLE QUERY 1

Use case: find all cardiac catheterization studies done in 2022

```
SELECT DISTINCT p.PERSON_ID, emd.PATIENT_ID, TO_CHAR(etd.TAG_VALUE) AS "StudyDate",  
               emd.STUDY_INSTANCE_UID, emd.SERIES_INSTANCE_UID, emd.IMAGE_FILE_PATH  
FROM CDMPHI.PROCEDURE_OCCURRENCE po  
  JOIN CDMPHI.PERSON p ON po.PERSON_ID = p.PERSON_ID  
  JOIN CDMECHO.ECHO_METADATA emd ON p.XTN_PATIENT_EPIC_MRN = emd.PATIENT_ID  
  JOIN CDMECHO.ECHO_TAGS_DATA etd ON emd.SOP_INSTANCE_UID = etd.SOP_INSTANCE_UID  
WHERE po.PROCEDURE_SOURCE_VALUE = 'CARDIAC CATH (CVI)'  
      AND etd.TAGS = 'StudyDate'  
      AND TO_CHAR(etd.TAG_VALUE) BETWEEN '20220101' and '20221231';
```

LOADING DICOM IMAGES

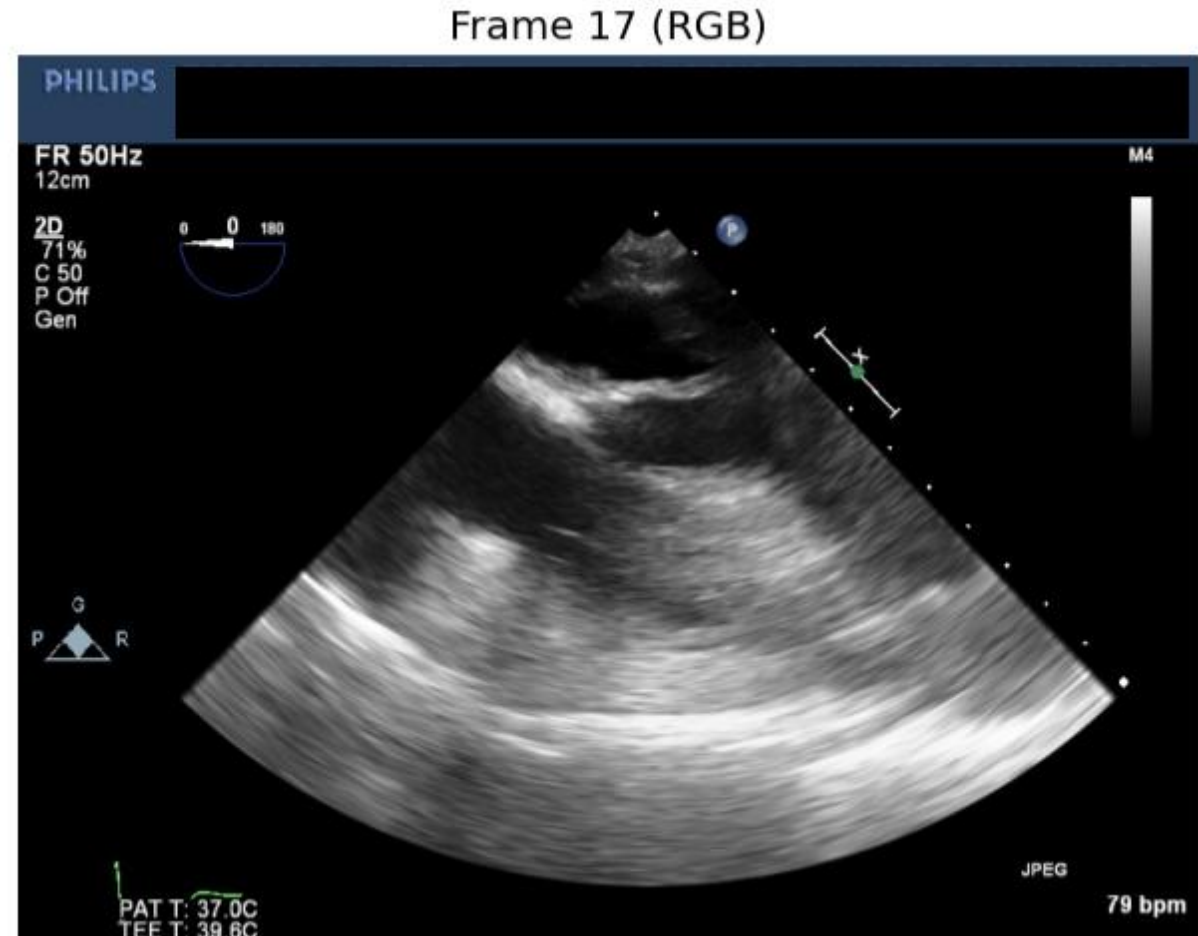
Example: read a DICOM file from an echocardiogram series, and display a frame in a notebook.

```
import pydicom
import matplotlib.pyplot as plt

ds = pydicom.dcmread("filename")
arr = ds.pixel_array

# pick a frame to show
frame_idx = arr.shape[0] // 2
frame = arr[frame_idx]

plt.figure(dpi=120)
plt.imshow(frame)
plt.title(f"Frame {frame_idx} (RGB)")
plt.axis("off")
plt.show()
```



Accessing the data

- PHI dataset – needs IRB approval
- Sailpoint workflow is under development

CDMECHO - Pitfalls and Dangers

- Link to EHR data is in progress
- Duplicate studies
- PatientIDs don't match Epic MRNs in older files
- The underlying EHR -> cardiology PACS link is based on MRN + study date (not even study time)

Summary

Summary & Conclusions

- AIR·MS represents a valuable, integrated dataset accessible through SQL.
- We are investigating additional datasets over the coming months.
- As AIR·MS matures, an app layer will be built out where reusable machine-learning pipelines can be implemented on clinical data.
- Building an ecosystem to accelerate clinical research and discovery, and we appreciate your input!