Enabling Scalable Multimodal AI Research in STARR: Integrating PHI-Scrubbed Imaging and Clinical Data

Hannah Morgan-Cooper[1,2], Joe Mesterhazy[1,2], P Desai[1,2]
[1] Stanford Health Care , [2]Stanford School of Medicine

1. Background

There is an increasing demand for multimodal data sources to advance AI and machine learning applications in healthcare, particularly those that integrate clinical and imaging data. While medical imaging plays a critical role in diagnosis and treatment, its use in large-scale observational research remains constrained by privacy concerns, technical challenges, and interoperability limitations. As part of the **STAnford Medicine Research Data Repository (STARR)**—a cloud-based platform managed by the Stanford Medicine Research Technology team that hosts a range of linked, analysis-ready clinical datasets on Google Cloud Platform – we implemented the OHDSI Medical Imaging Extension (1) and developed a dedicated PHI scrubbing pipeline for imaging data. This pipeline performs pixel-level redaction and metadata cleansing to produce deidentified, research-ready DICOM files. By linking these deidentified images with structured imaging metadata in the *image_occurrence* table, we enable privacy-preserving access to imaging data alongside EHR-derived clinical information. Furthermore, we enhanced the imaging metadata tables with modality-specific and study-level attributes to support more robust search capabilities, cohort discovery, and downstream AI/ML development. Together, these efforts establish a scalable framework for multimodal research at Stanford.

2. Methods

2.1 PHI-Scrubbed DICOM Files

Our goal was to produce PHI-scrubbed DICOM files that are structurally indistinguishable from the original raw files, enabling researchers to use them without making any changes to existing code or analysis pipelines. The scrubbing pipeline was designed to comprehensively address both pixel-level and metadata-based PHI risks while preserving data utility for downstream AI and research applications (Figure 1).

The PHI scrubbing process includes the following key components:

- **Pixel Redaction:** Automated detection and removal of burned-in identifiers from image pixels
- **Fallback Metadata Replacement:** Sensitive DICOM tags (e.g., *PatientID*, *AccessionNumber*) are replaced with standardized placeholders when PHI is detected.

- **Parameterized Tagging:** Researchers can optionally supply custom, user-defined values to replace specific tags, supporting flexible data labeling and study-specific needs.
- **Allowlist-Based Redaction:** Only explicitly permitted (safe) DICOM attributes are retained; all non-allowlisted tags are removed to minimize PHI risk.
- **UID Hashing:** DICOM Unique Identifiers (UIDs) are hashed using a consistent, non-reversible algorithm to preserve referential integrity across related files while preventing exposure of original identifiers.
- **Modality-Specific Rules:** Tailored scrubbing rules ensure relevant fields are preserved for each modality. For example, allowing modality specific known ‘PHI-safe’ metadata fields to pass through without redaction.
- **Regex-Based Redaction:** A two-stage regex-based pipeline identifies and masks dates, emails, numeric identifiers, and sensitive free-text, replacing unapproved tokens while preserving syntax.

All scrubbed DICOMs are stored in a secure GCP bucket with strict access controls. Each file retains valid DICOM formatting to ensure compatibility with existing research tools and pipelines.

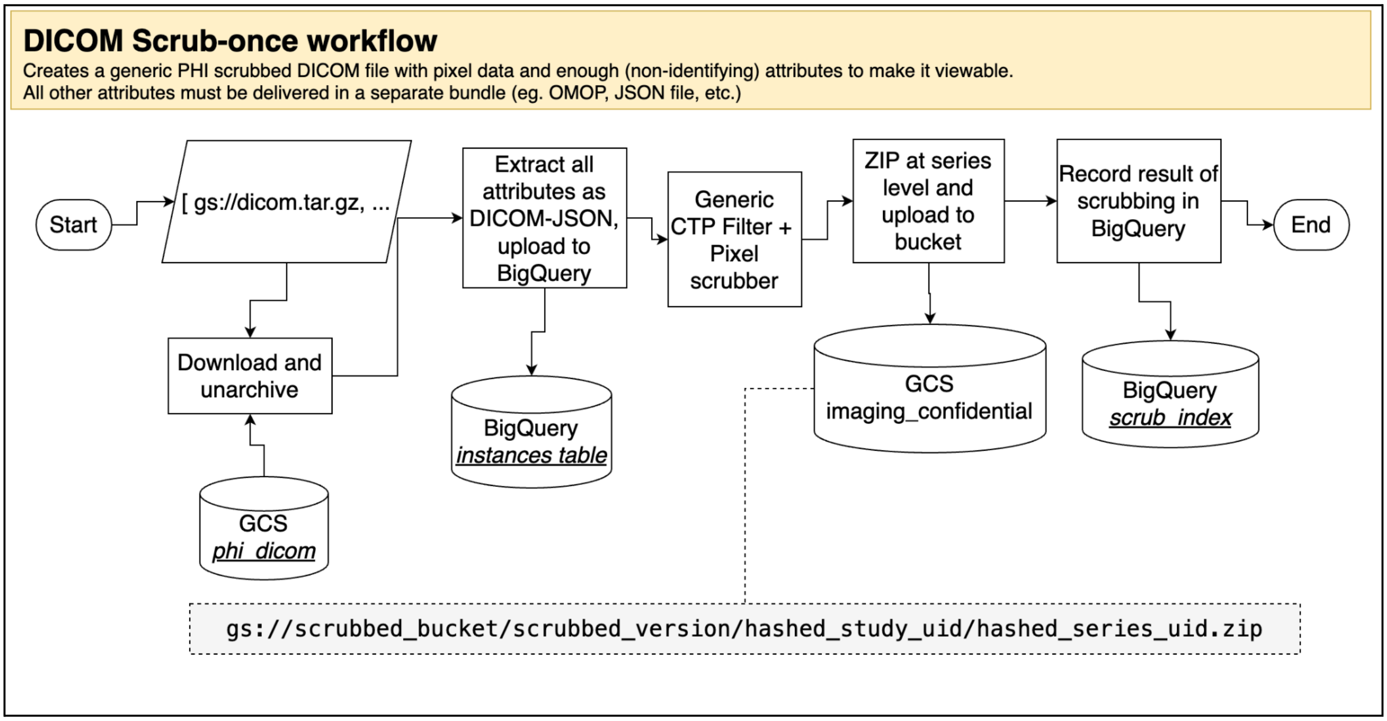
DICOM Scrub-once workflow
Creates a generic PHI scrubbed DICOM file with pixel data and enough (non-identifying) attributes to make it viewable.
All other attributes must be delivered in a separate bundle (eg. OMOP, JSON file, etc.)
Start
[gs://dicom.tar.gz, ...
Extract all attributes as DICOM-JSON, upload to BigQuery
Generic CTP Filter + Pixel scrubber
ZIP at series level and upload to bucket
Record result of scrubbing in BigQuery
End
Download and unarchive
GCS phi_dicom
BigQuery instances table
GCS imaging_confidential
BigQuery scrub_index
gs://scrubbed_bucket/scrubbed_version/hashed_study_uid/hashed_series_uid.zip

Figure 1: DICOM PHI scrubbing workflow

2.2 Implementation of the image_occurrence table

The image_occurrence table serves as the entry point for researchers to explore imaging data in OMOP. To facilitate faster cohort discovery and imaging study filtering, we added specific fields such as ***study_description, series_description, image_datetime, accession_number,***

and note_id. To reliably link imaging studies with EHR data, we implemented a composite key combining ***MRN, Date of Birth (DOB), and Accession Number***. This was necessary because accession numbers alone were not unique across different hospital departments. Furthermore, we applied PHI scrubbing to the **image_occurrence** table itself to support pre-IRB exploratory research and cohort building using de-identified OMOP data. Additional details are described in Table 1.

Table 1: PHI-scrubbed Image_occurrence table details

Field	**Type**	**Required**	**Description**	**phi scrubbing operation**
image_occurrence_id (PK)	INTEGER	Yes	The unique key that is given to an imaging study record.	Offset
person_id (FK)	INTEGER	Yes	A foreign key identifier to the person in the person table	Substitute
series_description	STRING	Yes	Series description as it appears in the DICOM metadata	Allow list redaction
study_description	STRING	Yes	Study description as it appears in the DICOM metadata	Allow list redaction
procedure_occurrence_id	INTEGER	Yes	A foreign key identifier to the procedure in the procedure table	Offset
visit_occurrence_id	INTEGER	No	A foreign key identifier to the visit in the visit_occurrence table	Offset
anatomic_site_source_value	STRING	No	Body_part_examined	Allow list
wadors_uri	STRING	No	Path to the location of the DICOM within the VNA	Del
local_path	STRING	No	Path to the PHI-scrubbed DICOM file within a GCP bucket. This is only populated for series that have been PHI scrubbed.	Replaced with path to scrubbed DICOM
image_occurrence_date	DATE	Yes	The date the study started, from the DICOM metadata	Jitter
image_occurrence_datetime	DATETIME	Yes	The datetime the study started, from the DICOM metadata	Jitter
image_study_uid	STRING	Yes	Unique identifier for the study	Hash

image_series_uid	STRING	Yes	Unique identifier for the series	Hash
modality_source_value	STRING	Yes	Type of device, process, or method that originally acquired or produced the data used to create the instances (images) in this series, as appears in the source DICOM metadata.	Allow list
_accession_number	STRING	No	Accession number for the series, note that this is not guaranteed to be unique and must be combined with person_id to link between tables	Hash
_note_id (STARR construct for linkability)	INTEGER	No	A foreign key identifier to the OMOP NOTE table. In cases where more than one note can be linked, the latest note date is used.	Offset
source_flag (STARR construct for traceability.	STRING	No	Indicates which clarity source (Stanford Health Care or Stanford Children's Health) was linked to the image series.	Pass as is
load_table_id (STARR construct for traceability)	STRING	No	Indicates which source table is the primary source for the table.	Pass as is
trace_id (STARR construct for traceability)	STRING	No	Contains the clarity order_proc_id and study, and series instance uids to identify which row in the source data the information came from.	Del

2.3 DICOM Metadata in BigQuery

We indexed over 35 million imaging series (CT, MR, XR, US) from Stanford's Vendor Neutral Archive(VNA). For each series, key metadata fields such as **modality**, **anatomic site**, **timestamps**, and relevant **identifiers** were extracted and stored at the series level in **BigQuery**. This metadata populates the **image_occurrence** table within the OMOP CDM, enabling researchers to perform federated queries and cohort discovery across imaging and clinical data—without requiring direct access to raw image files.

Results & Conclusion

STARR's implementation of the image_occurrance table paired with the phi scrubbed imaging data allows multimodal cohort building using structured queries over imaging and clinical data. Researchers can filter by modality, anatomy, and acquisition time, and link imaging to diagnoses, procedures, or radiology notes. Scrubbed DICOMs remain format-compatible and are securely accessible via GCP and can be easily accessed via secure GCP paths using standard tooling (e.g., gsutil) The image_occurrence table provides rich, de-identified metadata to support cohort discovery without exposing PHI, while links to the note table enable alignment with clinical interpretation. We believe that this scalable, privacy-preserving framework supports AI and observational research and can serve as a model for the broader OHDSI network.

References

1. Park WY, Jeon K, et al: Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension, J Imaging Inform Med;37(2), 2024
2. Park C, You SC, Jeon H, et al: Development and validation of the radiology common data model (R-CDM) for the international standardization of medical imaging data. Yonsei Med J 63:S74-S83, 2022
3. Kalokyri V, et al, MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes, JCO Clinical Cancer Informatics, 2023