2. Image Processing Pipeline Components (Industry-Level Detail)

In regulated medical AI systems (e.g., potential SaMD under MHRA guidance), the **image processing pipeline** is the critical first stage ensuring secure, compliant handling of sensitive data. It focuses on ingestion of clinical formats and rigorous anonymisation to comply with UK GDPR, Data Protection Act 2018, and NHS IG Toolkit requirements. No real patient data is stored or processed beyond temporary anonymised instances; all operations are logged for auditability.

This pipeline is implemented as modular Python code (e.g., in /backend/image_processing/), using standard libraries to ensure reproducibility and verifiability.

Key Technologies and Implementation

- **Primary Library**: pydicom (for DICOM parsing/editing) + numpy/opencv-python (for pixel-level operations).
- **Alternative Toolkits**: Optional integration with GDCM (via gdcm Python wrapper) or dcmtk (via subprocess) for advanced cases.
- **Support for Formats**:
  - Primary: DICOM (.dcm) – full support for multi-frame, enhanced DICOM, etc.
  - Secondary: NIfTI (.nii/.nii.gz) – for processed/segmented outputs from prior steps (using nibabel).
- **Secure Ingestion**:
  - Files uploaded via secure endpoint (e.g., Flask/FastAPI with HTTPS, authentication).
  - Validation: Check SOP Class UID, Transfer Syntax; reject invalid/non-medical files.
  - Temporary in-memory processing; no persistent storage of originals.

Automated Anonymisation Script/Tool

Core module: anonymizer.py – applies a configurable profile.

Best Practices Implemented (Aligned with Standards):

- Strictly follows **DICOM PS3.15 Security and System Management Profiles – Basic Application Level Confidentiality Profile** (2024 version):
  - Remove or replace all PHI (Protected Health Information) attributes (Option: Clean Pixel Data + Clean Recognizable Visual Features).
  - Key actions: Delete (Action D), Replace with dummy (Z/D), Empty (X), Keep if safe (K).
- Specific handling:
  - Remove/replace: Patient Name, Patient ID, Patient Birth Date, Institution Name/Address, Referring Physician, Accession Number, etc.
  - UID handling: Generate new Study/Series/SOP Instance UIDs (using pydicom.uid.generate_uid()).

- Date/Time: Shift all dates by a consistent random offset (e.g., +random days between 1-365) to preserve intervals but obscure absolutes.
- Retain clinical essentials: Modality, Series Description, Body Part Examined, Protocol Name (unless they contain PHI).

- **Burned-in Annotations Removal**:
    - Detect and blackout text/identifiers burned into pixels (common in secondary captures).
    - Method: Use OpenCV with OCR (Tesseract via pytesseract) to locate text regions, then apply black rectangles or inpainting.
    - Fallback: Heuristic masking of common overlay areas (top-left/right corners, bottom strips).

Industry-Level Example Diagrams and Visuals

Figure 14: TCIA-standard image de-identification and processing pipeline flowchart (Comprehensive end-to-end anonymisation flow – adapt for your ingestion + pixel cleaning.)

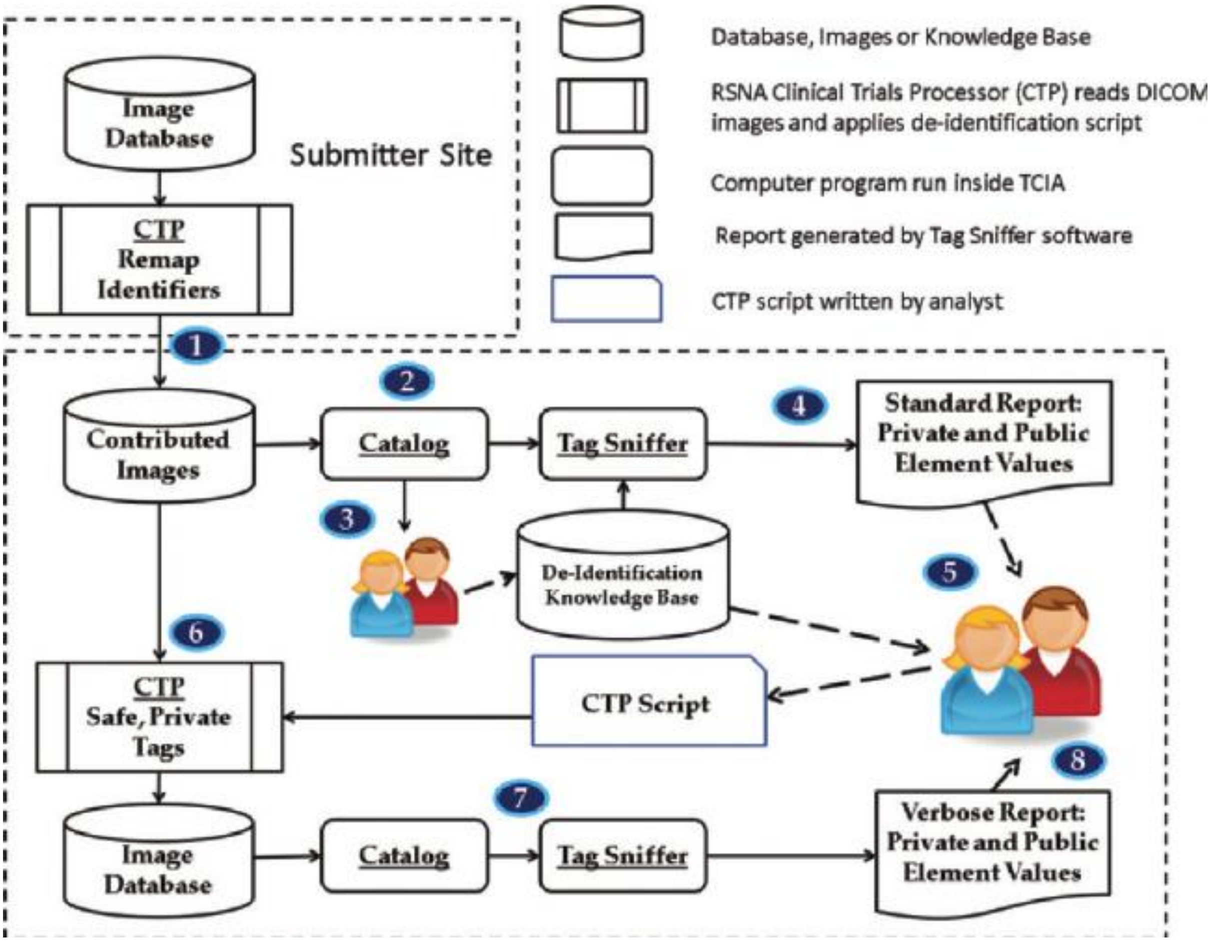
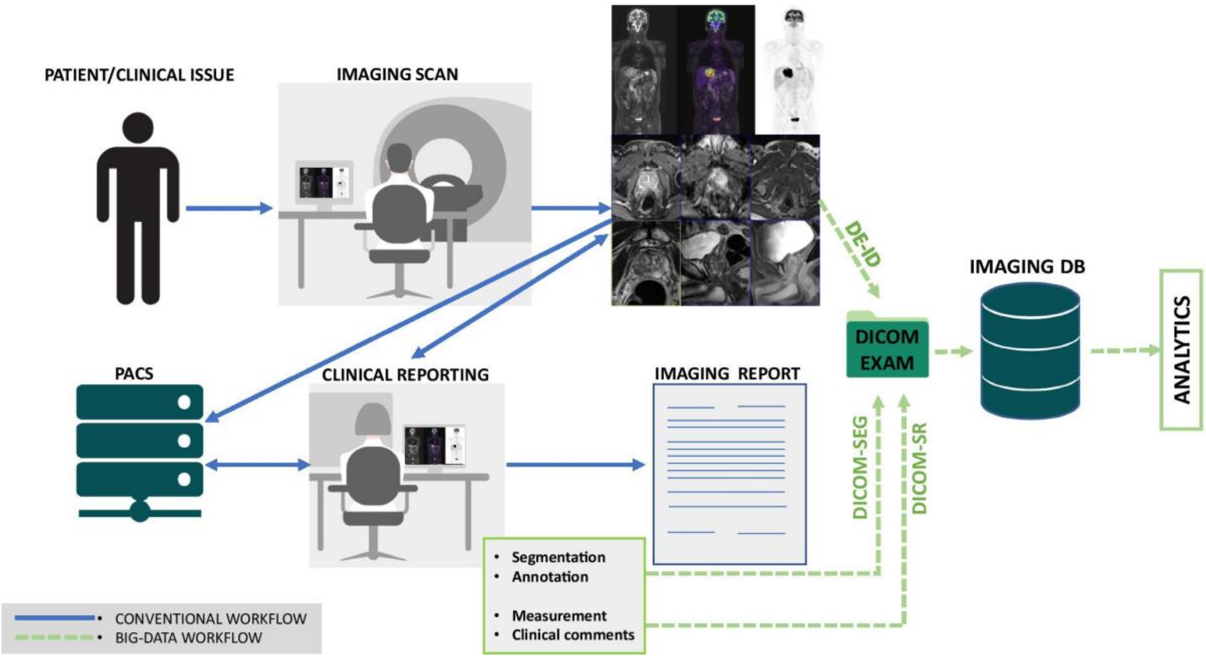


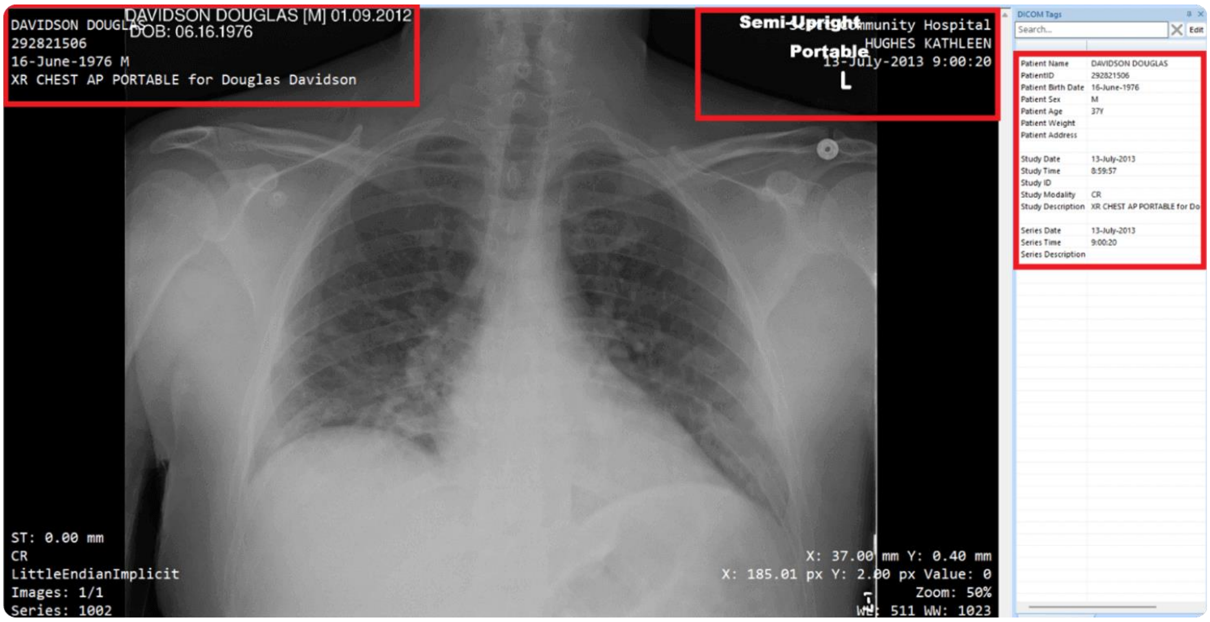
Image Database
CTP Remap Identifiers
Submitter Site
Database, Images or Knowledge Base
RSNA Clinical Trials Processor (CTP) reads DICOM images and applies de-identification script
Computer program run inside TCIA
Report generated by Tag Sniffer software
CTP script written by analyst
1
2
Contributed Images
Catalog
Tag Sniffer
4
Standard Report: Private and Public Element Values
3
De-Identification Knowledge Base
5
6
CTP Safe, Private Tags
CTP Script
8
7
Image Database
Catalog
Tag Sniffer
Verbose Report: Private and Public Element Values

researchgate.net

TCIA image de-identification process. Flowchart shows image ...

Figure 15: DICOM big data management and anonymisation pipeline (Shows ingestion, metadata handling, and de-identification steps.)





DAVIDSON DOUGLAS
292821506
16-June-1976 M
XR CHEST AP PORTABLE for Douglas Davidson

DAVIDSON DOUGLAS [M] 01.09.2012
DOB: 06.16.1976

Semi-Upright
Portable
L

Community Hospital
HUGHES KATHLEEN
13-July-2013 9:00:20

ST: 0.00 mm
CR
LittleEndianImplicit
Images: 1/1
Series: 1002

X: 37.00 mm Y: 0.40 mm
X: 185.01 px Y: 2.00 px Value: 0
Zoom: 50%
WL: 511 WW: 1023

DICOM Tags

Search...

Patient Name	DAVIDSON DOUGLAS
PatientID	292821506
Patient Birth Date	16-June-1976
Patient Sex	M
Patient Age	37Y
Patient Weight	
Patient Address	
Study Date	13-July-2013
Study Time	8:59:57
Study ID	
Study Modality	CR
Study Description	XR CHEST AP PORTABLE for Do
Series Date	13-July-2013
Series Time	9:00:20
Series Description	

opswat.com

DICOM Anonymization: Enhance Data Security with OPSWAT Proactive ...

Figure 19: Additional burned-in annotation cleaning example in medical images
(Shows region detection and masking.)

physicamedica.com

Data preparation for artificial intelligence in medical imaging: A ...

Evidence Artefacts for Repository

Place these in /docs/image_processing/ and /tests/:

- anonymizer.py: Full script with comments referencing PS3.15 tables (e.g., Annex E attributes).

- Sample logs: anonymization_log_sample.txt – showing processed tags, date shifts.
- Before/after headers: sample_before_header.txt and sample_after_header.txt (from public datasets like TCIA).
- Pixel cleaning demo: Screenshots or saved images showing original vs blacked-out regions.
- Test scripts: test_anonymizer.py (using public DICOM samples from pydicom datasets).

This pipeline ensures full traceability, minimal re-identification risk, and alignment with UK healthcare data governance. For assessors, include a README in the module explaining rationale and compliance mapping to PS3.15.