UNIVERSITY OF THE WESTERN CAPE
Private Bag X17, Bellville 7535, South Africa
Tel: +27 21 959 2911 | www.uwc.ac.za

Data Collection Protocol

Study BM25/3/12 – ML Healthcare Diagnostics

Principal Investigator:	Thabo Molefe
Supervisor:	Prof. Sarah van der Berg
Protocol Version:	2.0
Effective Date:	1 February 2026

1. Data Sources

Images will be obtained from two sources: (A) Tygerberg Hospital PACS archive - 8,000 posterior-anterior chest X-rays from adult patients (2020-2025), and (B) Groote Schuur Hospital PACS archive - 4,000 images from adults and paediatric patients (2021-2025).

2. Inclusion Criteria

PA chest X-rays of diagnostic quality; patients aged 12 and above; images with confirmed pathology status (GeneXpert-confirmed TB positive, or clinically confirmed negative); images taken with digital radiography equipment.

3. Exclusion Criteria

Lateral or AP views; severely degraded images (motion artifact, gross under/over-exposure); images from patients who have withdrawn consent; duplicate images from the same patient encounter.

4. Anonymisation Procedure

Step 1: Hospital radiology staff export images from PACS. Step 2: DICOM header stripping using PyDICOM library (removal of patient name, ID, DOB, all UIDs). Step 3: Assignment of random study ID (format: UWC-TB-XXXXX). Step 4: Transfer to research server via secure encrypted channel (SFTP). Step 5: Verification of anonymisation completeness by independent check.

5. Data Storage and Security

All data stored on UWC Research Computing Server (RCS). Access restricted to PI and two approved researchers via multi-factor authentication. Nightly encrypted backups. No data stored on personal devices, removable media, or cloud services. Server located in UWC data centre (physically secured, access-controlled).

6. Data Management Plan

Raw images stored in DICOM format. Processed images stored as PNG (512x512 pixels, 8-bit grayscale). Metadata stored in PostgreSQL database on same server. All processing scripts version-controlled in private Git repository. Data retention: 5 years post-study completion, then secure deletion with documented verification.

7. Quality Assurance

Random 5% sample audited monthly for anonymisation completeness. Image quality assessment using automated metrics (SNR, contrast). Inter-rater reliability testing for manual annotations (target kappa > 0.85). Protocol deviations logged and reported to BMREC within 7 days.