Clinical Dataset Curation Analysis

Binisha Thakur

Date: 03-02-2026

Dataset Discovery & Justification

DATASET 1: LIDC-IDRI (Lung Image Database Consortium)

Attribute	**Details**
Dataset Name	LIDC-IDRI | Data from The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans
Source Link	https://www.cancerimagingarchive.net/collection/lidc-idri/
Modality	Clinical Thoracic CT scan of Human Chest
Condition Covered	Lung Cancer, Non-Cancer (Benign), Metastatic Disease, Lung Nodules detection and characterization.

Why this dataset is useful for diagnostics:

- It is a web-accessible international resource for development, training, and evaluation of Computer-Assisted Diagnostic (CAD) methods for lung cancer detection and diagnosis.

- This dataset is a public-private partnership of National Cancer Institute (NCI), Foundation for the National Institutes of Health (FNIH) and Food and Drug Administration (FDA) which demonstrates the success of a consortium founded on a consensus-based process.

- This dataset contains 1018 cases each independently reviewed by four experienced thoracic radiologists in a rigorous two-phase process (blinded and unblinded), providing highly reliable ground truth for nodule identification.

- Nodule characterization is based on three categories:

  1. Nodule >= 3mm
  2. Nodule < 3mm
  3. Non-Nodule >= 3mm

- Semantic characterizations - Nine malignancy-related features (subtlety, spiculation, lobulation, etc.) enabling explainable AI.

- Not all nodules have biopsy-confirmed malignancy labels; malignancy scores are radiologist assessments, not true pathological ground truth.
- Definitive diagnosis data is available only for a subset of cases, limiting supervised cancer classification reliability.

Clinical Impact: AI trained on this dataset can serve as a second reader to reduce radiologists workload and catch overlooked nodules, potentially saving thousands of lives annually.

DATASET 2: PTB-XL (PhysioNet ECG Database)

Attribute	**Details**

Dataset Name	PTB-XL, a large publicly available electrocardiography dataset
Source Link	https://doi.org/10.13026/kfzx-aw45
Modality	Electrocardiography (ECG) – 12 Lead
Condition Covered	71 different cardiac abnormalities including myocardial infarction, conduction disturbances, hypertrophy, arrhythmias, and normal ECG patterns

Why this dataset is useful for diagnostics:

The PTB-XL dataset is critical for developing automated ECG interpretation systems, addressing a major global health need. Cardiovascular disease causes 17.9 million deaths annually, and ECG is the primary non-invasive diagnostic tools accessible even in primary care settings.

1. Comprehensive cardiac coverage: - 71 different diagnostic classes covering the full spectrum of cardiac abnormalities, from life-threatening arrythmias to chronic conditions.
2. Hierarchical Labelling: Labels organized into superclass (NORM, MI, STTC, CD, HYP) and subclasses, enabling both coarse and fine-grained diagnostic models.
3. Expert validation: Two cardiologists independently annotated each ECG, with consensus labels for training and individual labels for studying diagnostic variability.
4. Standardized Format: 12-lead ECGs at 500Hz , following international standards, ensuring compatibility with clinical equipment worldwide
5. Metadata richness: Includes patient age , sex, device information, and recording quality metrics.
6. Official stratified 10-fold splits are provided to prevent patient-level data leakage and ensure reproducible benchmarking.
7. Signal quality metadata (baseline drift, static noise, burst noise) enables automated quality filtering before model training.

Clinical Impact: AI trained on PTB-XL can provide instant preliminary ECG interpretation in emergency departments, rural clinics, and ambulance settings where cardiologists aren't immediately available. Automated arrhythmia detection can trigger alerts for life-threatening conditions (ventricular fibrillation, complete heart block) requiring immediate intervention. In low-resource settings, this democratizes access to expert-level cardiac diagnostics.

Authenticity & Clinical Reliability Assessment

Assessment Criteria	**LIDC-IDRI (CT-Lung Nodules)**	**PTB-XL (ECG – Cardiac Abnormalities)**
SOURCE CREDIBILITY		
Institution	National Cancer Institute (NCI) + Foundation for National Institute of Health (FNIH) + Food and Drug Administration (FDA)	Physikalisch-Technische Bundesanstalt (PTB) - German National Metrology Institute + Leipzig University Hospital
Journal/Platform	Primary: Medical Physics (2011); DOI: https://doi.org/10.1118/1.3528204; Supporting: SPIE Journal of Medical Imaging (2016); DOI: https://doi.org/10.1117/1.JMI.3.4.044504 / The Cancer Imaging Archive (TCIA)	PhysioNet (MIT-managed) - peer-reviewed in Scientific Data (Nature) 2020; DOI: https://doi.org/10.13026/6sec-a640

Credibility Score	Highest - Government research initiative	Highest - National metrology institute + peer-reviewed publication
DATASET SIZE AND DIVERSITY		
Sample Size	Total 1018 Patients, 1010 subjects considered (8 were scanned at 2 timepoints), 7371 lesions annotated	21,837 ECG Records from 18,885 unique patients
Diversity Strengths	Age range: 21-93 yrs Both screening and diagnostic CTs Multiple Scanner manufacturers Various Nodule Sizes (3-30mm)	Age Range: 0-95 yrs (full lifespan) Balanced gender: 52%male, 48%female Mix of impatient/outpatient recordings Multiple ECG devices 71 different diagnostics classes
Diversity Gaps	• Predominantly from one institution (limited geographic/ethnic diversity) • Likely US population bias	• European population (Germany) • Limited ethnic diversity documentation • Two institutions only (PTB + Leipzig) • Healthier population bias (outpatient clinic focus)
ANNOTATION METHOD		
Primary labelling	Expert-radiologists labelled (4 readers per case)	Two independent cardiologists per ECG
Quality Control	Consensus ground truth established Two phase annotation (nodule detection + characterization)	Consensus mechanism (third cardiologist resolves conflicts) Hierarchical SCP-ECG standard codes Quality validated against clinical reports Confidence level provided (0-100%)
Annotation Reliability	Very High – Gold Standard for nodule detection	Very High – Expert cardiologists consensus with confidence scores
Inter-rate Agreement	Published: k = 0.48-0.69 for semantic features	Published: Agreement varies by diagnostic class (higher for MI, lower for subtle ST changes)
ETHICAL CONSIDERATIONS		
Patient Consent	Institutional review board approved Public research use permitted	Ethics Committee approved (Charite ethics Committee) Patients provided informed consent
Anonymization	Full DICOM de-identification PHI removed per HIPAA standards	De – identified per EU GDPR standards Patient metadata removed except age/sex

Data Use Agreement	Required – must register and agree to terms	Open Data Commons Open Database License (ODbL) Attribution required
Privacy Risk	Low – robust de - identification	Low – GDPR – compliant anonymization
POTENTIAL BIASES and GAPS		
Population Bias	Geographical: US-Centric Likely underrepresents minority groups Socioeconomic Bias (insured patients)	Geographic: European (German) population Maay not generalize to populations with different baseline cardiovascular risk profiles Predominantly Caucasian ethinicity (assumed)
Clinical Bias	Enriched for abnormal findings (screening + diagnostic mix) May not reflect true screening prevalence	Outpatient clinic bias (less acute/severe cases than ER data) Disease prevalence: Only 9% normal ECGs (enriched for abnormalities) Underrepresents emergency arrhythmias
Technical Bias	Mix of slice thicknesses (0.6-5mm) Different reconstruction kernels	Multiple ECG devices (Schiller AT-1, AT-10, etc.) Some records at 100 Hz (down sampled) vs native 500 Hz Baseline wander and noise levels vary
Missing Data	Minimal - comprehensive metadata	Some ECGs lack detailed clinical context (medications, symptoms) 15% of records have quality issues flagged
CLINICAL RELIABILITY RATING		
Suitability for Training	Excellent for supervised learning	Excellent - large scale + expert labels + hierarchical structure
Clinical Deployment Readiness	Requires external validation on diverse populations before deployment	Requires validation across different populations (ethnicity, geography) and acute care settings (ER, ICU) before broad deployment
KEY STRENGTHS	Gold-standard annotations Semantic features enable explainable AI multi-reader consensus	Large scale (21K+ ECGs) Hierarchical labels (flexibility) Confidence scores (uncertainty quantification) High sampling rate (500 Hz) Balanced gender distribution
KEY LIMITATIONS	Limited diversity Smaller sample size vs. modern datasets Expensive annotation approach (not scalable)	Geographic/ethnic homogeneity Outpatient bias (less severe cases) Imbalanced class distribution (9% normal)

		No clinical outcomes data 15% have quality flags

Labelling Framework Design

Dataset: LIDC-IDRI (Lung Image Database Consortium)
Modality: CT (Computed Tomography)
Date: 03/02/2026

Label Category	Type	Values	Clinical Significance
Nodule Classification	Multi-Class	Nodule>=3, Nodule < 3mm, Non-nodule	Defines inclusion in analysis
Subtlety	Ordinal	1-5 scale	Detection difficulty
Internal Structure	Categorical	1-Solid, 2-Fluid, 3-Fat, 4-Air	Tissue composition
Calcification	Categorical	1-Popcorn, 2-Laminated, 3-Solid, 4-Non-central, 5-Central, 6-Absent	Benign vs malignant indicator
Sphericity	Ordinal	1-5 scale	Shape regularity
Margin	Ordinal	1-5 scale	Edge definition
Lobulation	Ordinal	1-5 scale	Surface irregularity
Spiculation	Ordinal	1-5 scale	Spiky projections (high malignancy risk)
Texture	Categorical	1-GGO, 2-Part-solid, 3-Solid	Density pattern
Malignancy	Ordinal	1-5 scale	Overall cancer likelihood

Metadata Fields:

- Patient_ID, Series_UID, Nodule_ID
- Annotator_ID, Date_Annotated
- Slice_Start, Slice_End, Max_Diameter_mm, Volume_mm3
- Image_Quality (1-5), Notes

Inclusion & exclusion criteria

INCLUDE:

- ✓ All nodules ≥3mm with complete semantic characterization
- ✓ Small nodules <3mm (basic detection label only)
- ✓ non-nodule structures ≥3mm (for false positive training)
- ✓ multi-reader annotations (captures inter-observer variability)
- ✓ Scans with slice thickness <5mm

EXCLUDE:

- ✗ Severe motion artifacts (non-diagnostic)
- ✗ Missing DICOM metadata
- ✗ Corrupted/incomplete series
- ✗ Duplicate scans from same patient

Rationale: Ensures dataset quality while maintaining real-world variability needed for robust AI training.

Consistency assurance protocol

A. Standardized Annotation Environment

- **Software:** 3D Slicer with standardized lung window settings (Level: -600, Window: 1500)
- **Review Protocol:** All three planes (axial, coronal, sagittal) + 3D reconstruction
- **Template:** Google Sheets with data validation rules enforcing valid ranges

B. Three-Tier Quality Control

Tier 1 - Automated Validation:

- Data validation rules prevent out-of-range entries (e.g., Subtlety must be 1-5)
- Automatic flags for inconsistencies (e.g., Malignancy=5 but Spiculation=1)

Tier 2 - Training & Calibration:

- All annotators label 20 practice cases independently
- Consensus meeting to resolve discrepancies ≥2 points
- Re-calibration until inter-rater agreement κ > 0.70
- In addition to consensus voting, probabilistic label aggregation (e.g., soft labels based on reader agreement frequency) may be used to preserve diagnostic uncertainty rather than forcing hard consensus.

Tier 3 - Senior Review:

- Random 10% of cases reviewed weekly
- 100% review of high malignancy ratings (4-5)
- Consensus panel for cases with ≥2-point disagreement

C. Rating Scale Definitions (Key Examples)

Subtlety:

- 1 = Extremely subtle (requires comparison with adjacent slices)
- 3 = Fairly subtle (visible with careful examination)
- 5 = Obvious (immediately visible)

Spiculation:

- 1 = None (smooth edges)
- 3 = Moderate (several spikes visible)
- 5 = Marked (corona radiata sign, extensive spiculation)

Malignancy:

- 1 = Highly unlikely (calcified, stable)
- 3 = Indeterminate (needs follow-up)
- 5 = Highly suspicious (multiple malignant features, recommend biopsy)

Implementation approach

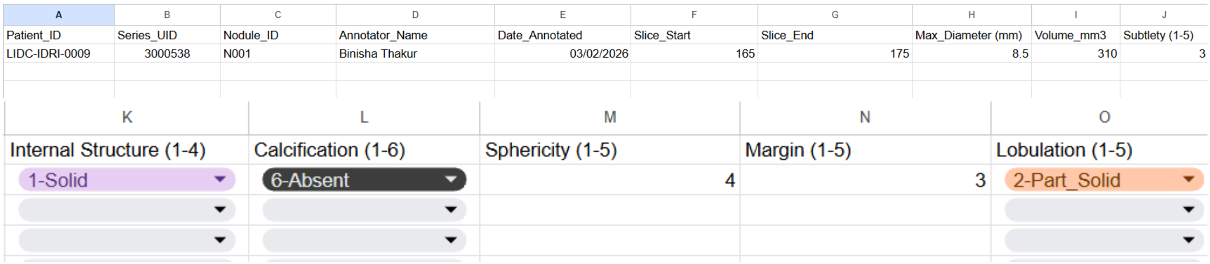
Platform Justification:

- **3D Slicer:** Free, open-source DICOM viewer with multiplanar reconstruction and measurement tools
- **Google Sheets:** Real-time collaboration, built-in data validation, version control, audit trail

Workflow Per Case:

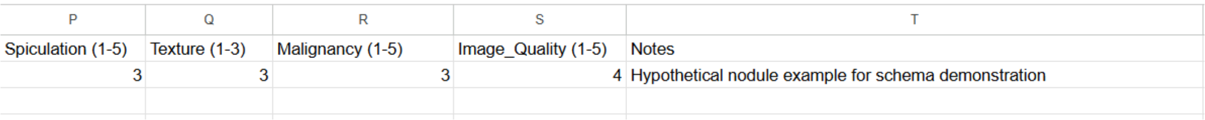
1. Load CT in 3D Slicer with lung window preset
2. Identify all nodules ≥3mm
3. For each nodule: measure diameter, rate 9 semantic features
4. Document immediately in validated spreadsheet
5. Quality checklist before saving

Screenshot: Annotation Template with Data Validation

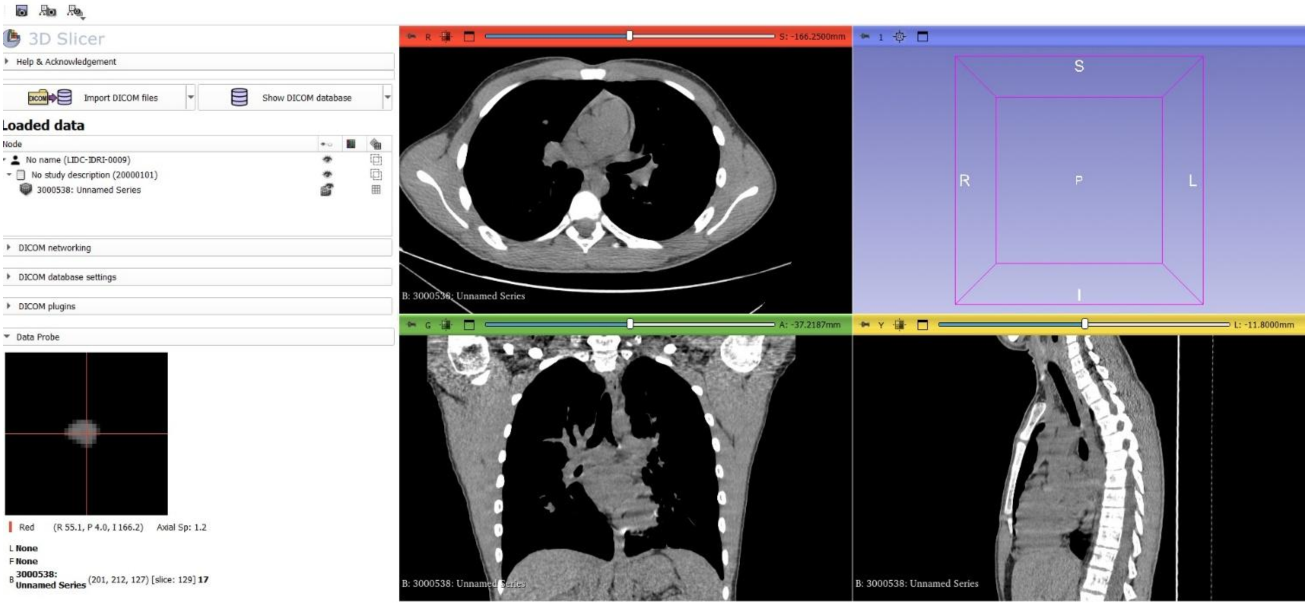


A	B	C	D	E	F	G	H	I	J
Patient_ID	Series_UID	Nodule_ID	Annotator_Name	Date_Annotated	Slice_Start	Slice_End	Max_Diameter (mm)	Volume_mm3	Subtlety (1-5)
LIDC-IDRI-0009	3000538	N001	Binisha Thakur	03/02/2026	165	175	8.5	310	3

K	L	M	N	O
Internal Structure (1-4)	Calcification (1-6)	Sphericity (1-5)	Margin (1-5)	Lobulation (1-5)
1-Solid	6-Absent	4	3	2-Part_Solid



P	Q	R	S	T
Spiculation (1-5)	Texture (1-3)	Malignancy (1-5)	Image_Quality (1-5)	Notes
3	3	3	4	Hypothetical nodule example for schema demonstration



Data Filtering & Quality Control Plan

Dataset: PTB-XL ECG Database

STEP 1: Remove Poor-Quality or Irrelevant Signals

Automated Technical Checks:

Checks	Criteria	Action
File integrity	All 12 leads present, 10-second duration	Reject if incomplete
Sampling rate	500 Hz (or 100 Hz acceptable)	Keep both, flag 100 Hz for certain analyses
Signal-to-Noise Ratio	SNR > 15 dB per lead	Reject if <15 dB in ≥3 leads
Baseline wander	Deviation < 0.5 mV	Apply high-pass filter; reject if unfixable
Artifacts	Power line noise, muscle artifacts, clipping	Apply filters; reject if severe
PTB-XL quality flags	Check: baseline_drift, static_noise, burst_noise	Reject if burst_noise or severe flags present

Manual Review:

- Random 5% sample reviewed by clinical expert
- Verify automated decisions are correct
- Adjust thresholds if needed

Expected Outcome: Reject ~5% of records due to quality issues

STEP 2: Handle Missing or Ambiguous Labels

Decision Framework:

Issue	Resolution
Low confidence (<50%)	Exclude from training; use in test set only (challenging cases
Medium confidence (50-75%)	Include but with lower weight in loss function
High confidence (>75%)	Full weight in training
Inter-cardiologist disagreement	If different superclass → Use third cardiologist consensus If same superclass → Use agreed superclass label
Missing age/sex	Age missing → Reject (critical for diagnosis) Sex missing → Keep but exclude from sex-stratified analysis
Rare classes (<10 examples)	Exclude entire class (insufficient for training)

STEP 3: Deal with Class Imbalance

Current Distribution Problem:

NORM (Normal):	9% ← MINORITY
STTC (ST/T Change):	24% ← MAJORITY
MI, CD, HYP:	5-16% each (approx.)

Balancing Strategy:

Approach	Implementation
Stratified Split	Train/Val/Test maintain class ratios **Critical:** Split by patient, not ECG (prevent data leakage)
Class Weights	Weight = 1/√(class_frequency) Penalize misclassifying rare classes more
Oversample NORM	Duplicate or augment normal ECGs to reach ~15% in training
Focal Loss	Focus learning on hard-to-classify examples
Safe Augmentation	Only for minority classes: • Amplitude scaling (±10%) • Baseline shift (±0.2 mV) • Time stretch (±2%) • Light Gaussian noise Never: large time warping, lead swapping
No Under sampling	Keep all data; don't throw away majority class examples

STEP 4: Usability Decision - "Accept" or "Reject"

Final Decision Tree:

For each ECG:

Technical Quality, OK?

NO (corrupt, incomplete, SNR<15) → REJECT
YES → Continue

Has valid diagnostic label?
NO → REJECT
YES → Continue

Confidence level?

<50% → TEST SET ONLY (challenging cases)
50-75% → TRAINING (lower weight)
>75% → TRAINING (full weight)

Age metadata present?

NO → REJECT
YES → Continue

Class has ≥10 examples?

NO → EXCLUDE class
YES → ACCEPT

QUALITY CONTROL SUMMARY

Input: 21,837 ECG records

After Filtering:

- **Usable:** ~20,650 (94.6%)
  - Gold Standard: ~18,400 (84%)
  - Usable with lower weight: ~2,250 (10%)
- **Challenging (test only):** ~600 (2.7%)
- **Rejected:** ~600 (2.7%)

Key Principles:

1. **Automate where possible** - Technical checks, SNR calculation
2. **Expert review for ambiguity** - Low confidence labels, disagreements
3. **Patient-level splitting** - Prevent data leakage (same patient in train +x`x test)
4. **Keep challenging cases** - Use for model stress-testing, not training
5. **Document everything** - Log all rejection reasons for audit trail

Insight & Reflection

What is the biggest risk of using poorly curated diagnostic data?

The most critical risk is systematic misdiagnosis at scale, which directly violates the fundamental medical principle of "first, do no harm." When AI models are trained on mislabelled, biased, or low-quality data, they don't make random errors—they learn and perpetuate harmful patterns that can affect thousands of patients.

Specific Risks:

1. Population Bias and Healthcare Disparities
When training data overrepresents certain demographics, AI systems underperform on others, missing diagnoses in underrepresented populations and worsening existing healthcare inequalities across regions and ethnic groups.

2. False Negatives in Critical Conditions
Inaccurate or weak labels can train AI models to miss life-threatening conditions like arrhythmias or early-stage cancer, systematically repeating dangerous diagnostic errors across large patient populations.

3. Illusion of High Accuracy
Validation on flawed or biased test datasets can produce misleadingly high accuracy metrics, creating false confidence and resulting in unsafe real-world deployment with significant diagnostic failures.

4. Loss of Clinical Trust
Visible AI diagnostic errors reduce healthcare providers' trust in decision-support systems, slowing adoption of reliable tools and limiting the long-term benefits of AI integration in clinical practice.

5. Public Health Consequences
In resource-limited regions relying heavily on AI diagnostics, biased or inaccurate models can delay treatment, miss outbreaks, and amplify systemic healthcare gaps, impacting entire communities.

If given 6 months, how would I improve this dataset for clinical-grade AI?

MONTHS 1–2: DIVERSITY EXPANSION

1. Geographic Diversity Partnerships
Collaborate with hospitals across India, Southeast Asia, Africa, and South America to ensure dataset representation across varied ethnic, clinical, and healthcare infrastructure settings globally.

2. Demographic Representation Targets
Collect 500 additional cases with balanced age groups, equal gender distribution, documented ethnicity, and inclusion of both urban and rural socioeconomic backgrounds.

3. Equipment Diversity Inclusion
Incorporate scans from multiple manufacturers and older imaging systems to reflect real-world variability, especially in resource-limited healthcare environments.

4. Mandatory Clinical Metadata Collection
Standardize collection of age, sex, ethnicity, BMI, smoking history, comorbidities, and geographic location to enable subgroup performance analysis and bias detection.

5. Impact of Diversity Expansion
Training AI on globally representative data improves generalization, reduces algorithmic bias, and ensures safer deployment across heterogeneous patient populations.

MONTHS 3–4: EXPERT LABEL REFINEMENT & OUTCOME DATA

6. Longitudinal Outcome Linking
Connect CT scans with biopsy results, pathology reports, growth rates, and survival outcomes to replace subjective malignancy ratings with objective ground-truth evidence.

7. Ground-Truth Target Threshold
Secure follow-up outcomes for at least 30% of high-risk nodules, transforming weakly supervised labels into clinically verified diagnostic references.

8. Expert Re-annotation Initiative
Engage fellowship-trained thoracic radiologists to re-annotate 1,000 cases using consensus protocols and standardized evaluation criteria.

9. Pixel-Level Segmentation Enhancement
Add detailed segmentation masks for nodules ≥6mm, enabling precise localization and improving model learning beyond coarse bounding measurements.

10. Active Learning Prioritization
Focus expert review on cases where AI demonstrates highest uncertainty, maximizing annotation efficiency and strengthening model performance in ambiguous scenarios.

11. Imaging Acquisition Standardization
Implement consistent CT acquisition protocols and retrospectively filter substandard scans to reduce technical variability affecting model accuracy.

12. Importance of Ground-Truth Validation
Biopsy-confirmed outcomes represent the diagnostic gold standard, surpassing subjective expert opinion and ensuring clinically reliable model training.

MONTHS 5–6: PROSPECTIVE CLINICAL VALIDATION

13. Multi-Site Prospective Trial Deployment
Deploy the AI model across three international hospitals, comparing radiologist-alone performance versus radiologist-assisted AI workflows in real clinical settings.

14. Real-World Performance Measurement
Evaluate sensitivity, specificity, false positives, missed cancers, and time-to-diagnosis impact within actual screening populations.

15. Clinician Workflow Feedback Loop
Gather structured radiologist feedback to assess usability, workflow efficiency, trust levels, and identify scenarios where AI support is beneficial or problematic.

16. Failure Mode Documentation
Systematically analyze disagreements between AI and clinicians to understand edge cases and refine future model iterations.

17. Clinical Challenge Benchmark Creation
Develop a 200-case stress-test dataset featuring subtle, ambiguous, or difficult nodules to benchmark robustness in complex diagnostic scenarios.

18. Regulatory Documentation Preparation
Compile subgroup performance data, validation metrics, failure analyses, and demonstrated clinical utility for FDA 510(k) or CE Mark submission readiness.

DELIVERABLES AFTER 6 MONTHS

19. Enhanced Dataset Quality
Deliver 500 diverse cases, biopsy-linked ground truth for 30% of nodules, detailed segmentation masks, and documented diversity metrics.

20. Clinical Validation Evidence
Provide prospective multi-site trial results, subgroup-stratified performance analysis, challenge set benchmarking, and radiologist trust assessments.

21. Regulatory-Ready Submission Package
Prepare comprehensive documentation demonstrating safety, efficacy, subgroup fairness, clinical benefit, and clearly defined limitations for regulatory approval.

THANK YOU