

Predicting Pathological Complete Response to Neoadjuvant Chemotherapy in Breast Cancer Using High-Throughput Histopathological Images and Machine Learning

A Retrospective Study Protocol

AI-pCR

Department of Breast Oncosurgery, Tata Medical Center (TMC), Kolkata

Study duration: 3 years

Study data period: January 2011 – December 2025

Protocol Version: Version 1.1 | Date: 6th June 2026

PRINCIPAL INVESTIGATOR:

Dr. Sanjit Kumar Agarwal

CO-INVESTIGATORS:

Dr. Rosina Ahmed

Dr. Vidita Modi

Dr. Indu Arun

Saipradeep V.G.

Thomas Joseph

Glossary and Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BC	Breast Cancer
DCA	Decision Curve Analysis
H&E	Hematoxylin and Eosin
HER2	Human Epidermal Growth Factor Receptor 2
ML	Machine Learning
NACT	Neoadjuvant Chemotherapy
pCR	Pathological Complete Response
REDCap	Research Electronic Data Capture
TNBC	Triple-Negative Breast Cancer
TMC	Tata Medical Center
WSI	Whole-Slide Image

1) Rationale

Neoadjuvant chemotherapy (NACT) has become a cornerstone in the management of locally advanced and high-risk early breast cancer, particularly in triple-negative breast cancer (TNBC) subtype. By improving surgical feasibility and enabling breast and axillary conservation, NACT also provides critical in vivo information about tumour chemosensitivity. Pathological complete response (pCR), defined as the absence of residual invasive tumour in the breast and axilla at the time of surgery (ypT0/Tis ypN0 per AJCC staging), as assessed on the surgical resection specimen, is a well-established surrogate endpoint for long-term outcomes, including event-free survival and overall survival [1,2,3].

Despite its prognostic significance, pCR rates vary widely — ranging from 10% to 60% — depending on molecular subtype, treatment regimen, and individual patient characteristics [3,4]. In India, NACT decisions are guided primarily by clinical stage and tumour biology, yet no reliable, validated tools exist to predict individual response before treatment initiation. This gap results in a significant proportion of patients being exposed to the full toxicity of chemotherapy without achieving pCR, while others who might benefit from intensified regimens are undertreated.

Accurate pre-NACT pCR prediction could transform clinical decision-making: enabling personalized treatment selection, sparing patients from unnecessary toxicity, optimizing surgical planning, and reducing healthcare costs. Emerging evidence suggests that morphological features of the pre-treatment tumour microenvironment — including tumour-infiltrating lymphocytes (TILs), nuclear pleomorphism, mitotic activity, and stromal composition — are strongly associated with chemotherapy response and can be quantified from routine H&E-stained core-biopsy slides [5,6].

Recent advances in computational pathology and machine learning have made it increasingly feasible to extract clinically relevant signals from histopathology images. Broad deep learning breakthroughs in oncology and medical imaging have been followed by strong pathology-specific applications, including lymph node metastasis detection and cancer subtyping from tissue sections [9,10]. More recently, weakly supervised and data-efficient approaches have enabled model development from slide-level labels rather than exhaustive annotations, improving scalability for whole-slide images [13,14]. In parallel, transformer architectures have reshaped how large image contexts are modeled, supporting both multimodal co-attention strategies and foundation models for computational pathology [7,8,15]. Because robust translation requires careful attention to validation, generalizability, and clinical utility, this work is also informed by established guidance on study design in cancer diagnostics and computational histopathology [11,12].

This study aims to develop an indigenous, affordable, and interpretable AI-based model tailored to the Indian population for predicting pCR following NACT in TNBC, using pre-treatment H&E core-biopsy WSIs integrated with routinely collected clinical covariates from Tata Medical Center (TMC), Kolkata with a prospectively maintained database spanning 2011–2025, managed in REDCap for structured data capture and analysis. Model performance will be evaluated not only by discrimination and calibration, but also by decision curve analysis to assess potential clinical usefulness across threshold probabilities.

2) Methods

a. Study Design

- Study Type: Retrospective analysis of prospectively collected data
- Purpose: AI model development and internal validation
- Study Site: Tata Medical Center (TMC), Kolkata
- Duration: 3 Years (for model development and internal validation) for patients who underwent treatment at TMC from 2011 from 2025

b. Aims and Objectives

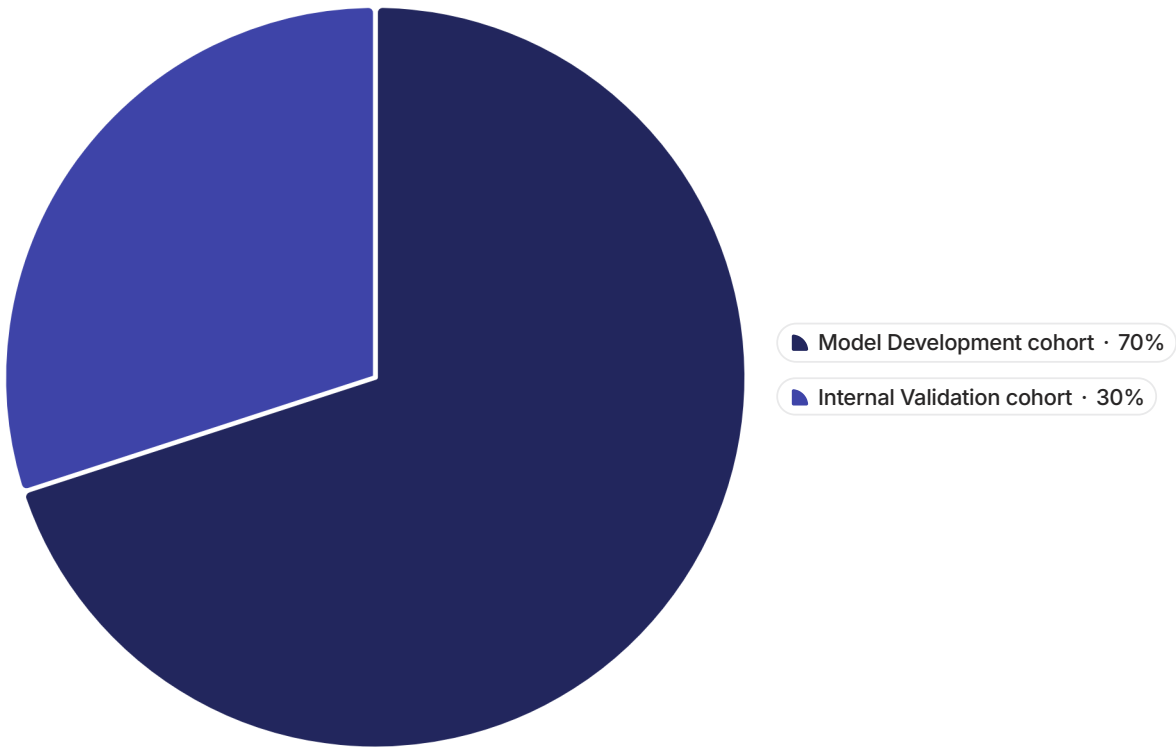
AIM: To develop an indigenous, affordable, and interpretable AI-based model tailored to the Indian population for predicting pathological complete response (pCR) following neoadjuvant chemotherapy (NACT) in triple-negative breast cancer (TNBC) breast cancer.		
OBJECTIVES:		
01	02	03
Digitize and Curate WSIs To digitize and curate pre-NACT H&E core-biopsy whole-slide images (WSIs) from TNBC breast cancer patients treated at TMC (2011–2025)	Extract Morphological Features To extract interpretable morphological and cellular features from WSIs using deep phenotyping pipelines (tumour-region localization, nuclear/cell-type composition, mitotic activity surrogates)	Develop AI Model To develop and validate a transformer-based AI model integrating WSI features and clinical covariates (age, stage, grade, receptor status) to predict pCR
04	05	
Evaluate Model Performance To evaluate model performance using nested cross-validation and assess calibration, discrimination (AUC-ROC), and clinical utility (decision curve analysis)	Identify Biomarkers To identify interpretable morphological biomarkers associated with pCR in the Indian breast cancer population	

c. Inclusion and Exclusion Criteria

Inclusion Criteria: 1. Age 18–70 years 2. T1–T3, N0–N1, M0 breast cancer 3. Core biopsy–proven Triple–Negative Breast Cancer (TNBC) breast cancer 4. Received NACT followed by definitive breast and axillary surgery at TMC 5. Treated at TMC between 2011–2025 6. Available pre-NACT H&E core-biopsy whole-slide images (WSIs) 7. Complete clinical data available in REDCap database	Exclusion Criteria: 1. Age <18 or >70 years 2. Non-TNBC molecular subtypes 3. M1 (metastatic) disease 4. Incomplete NACT course or no definitive surgery 5. Missing or inadequate pre-NACT pathology slides 6. Prior breast cancer treatment or neoadjuvant therapy 7. Incomplete clinical records in REDCap
--	---

d. Sample Size and Data Distribution

A total of 500 patients will be included:



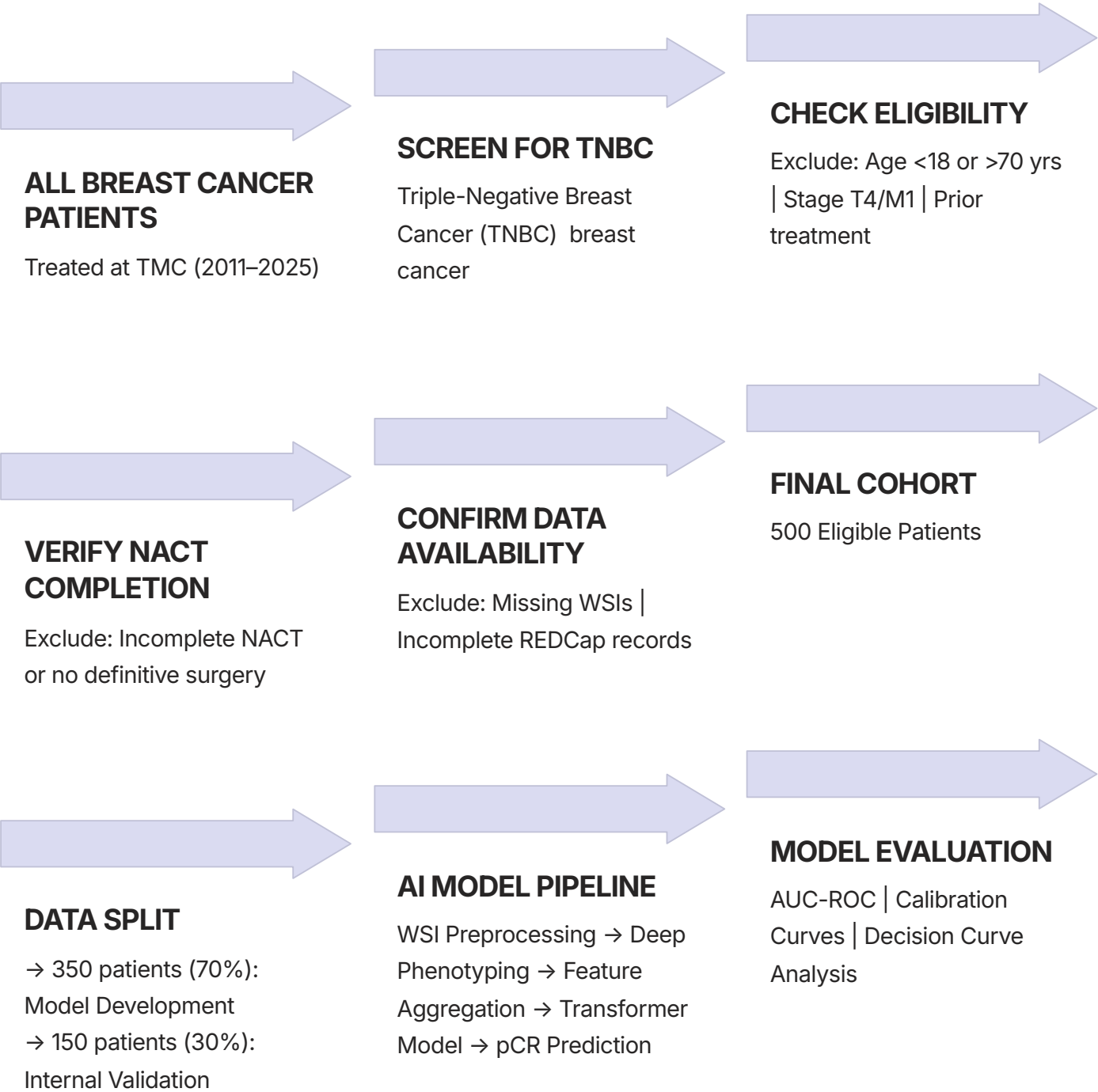
Data Sources: Departmental REDCap database (clinical data); Pathology laboratory (H&E slides)

e. Study Duration

3 years (model development and internal validation phase)

Study Flowchart

The following flowchart outlines the patient selection and data flow for AI model development and internal validation.



3) Data Management

Clinical data will be retrieved from the departmental REDCap (Research Electronic Data Capture) database maintained at Tata Medical Center (TMC). REDCap is a secure, web-based application designed to support data capture for research studies, providing an intuitive interface for validated data entry, audit trails for tracking data manipulation and export procedures, automated export procedures for seamless data downloads to common statistical packages, and procedures for data integration and interoperability with external sources.

Pre-NACT H&E-stained core-biopsy slides will be retrieved from the TMC pathology laboratory archive. Slides will be digitized into whole-slide images (WSIs) using a validated digital scanner at a minimum resolution of 20× magnification. All WSIs will undergo quality control assessment prior to inclusion in the analysis pipeline.

Data will be stored on secure institutional servers with restricted access. Patient identifiers will be anonymized prior to computational analysis. All data handling will comply with institutional data protection policies and applicable national regulations. A unique study ID will be assigned to each patient to maintain traceability while ensuring confidentiality.

The AI model development pipeline, including feature extraction and model training, will be conducted using de-identified data only. All code and model weights will be version-controlled and archived for reproducibility.

4) Statistical Considerations

a. Sample Size Consideration

This retrospective study is designed for AI model development and internal validation. A total of 500 patients (350 for model development, 150 for internal validation) has been determined based on the following considerations:

The expected pCR rate in TNBC breast cancer at TMC is approximately 30–40%, based on institutional data and published literature [1,2]. For a binary classification model (pCR yes/no), a minimum of 10 events per predictor variable is recommended for logistic regression-based approaches. For deep learning models applied to WSIs, a minimum of 200–300 cases per class is considered adequate for initial model development with transfer learning approaches [7,8,13,14].

Given an expected pCR rate of ~35%, a cohort of 500 patients would yield approximately 175 pCR-positive and 325 pCR-negative cases, providing sufficient class representation for model training and internal validation. The 70:30 split (350 development : 150 validation) ensures adequate sample sizes in both cohorts.

b. Planned Statistical Analyses

The primary performance metric for the AI model will be the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) [1,2]. Secondary metrics will include:

- Sensitivity and specificity at pre-specified operating thresholds
- Positive predictive value (PPV) and negative predictive value (NPV)
- Calibration assessed using calibration curves and Brier score
- Clinical utility assessed using decision curve analysis (DCA) [16]

Model development will employ nested cross-validation (outer loop: 5-fold; inner loop: 3-fold) [11] on the development cohort (n=350) for hyperparameter tuning and unbiased performance estimation. Final model performance will be reported on the held-out internal validation cohort (n=150).

Class imbalance will be addressed using oversampling techniques (SMOTE) or class-weighted loss functions during model training. Patient-level data separation will be strictly enforced to prevent data leakage between training and validation sets.

All statistical analyses will be performed using Python (scikit-learn, PyTorch) and R. A two-sided p-value of < 0.05 will be considered statistically significant for secondary analyses. Results of secondary outcome analyses will be interpreted as exploratory.

5) Safety

This is a retrospective study. No experimental treatment, diagnostic procedure, or intervention will be performed on patients as part of this study. All clinical data and pathological specimens used in this study were collected as part of routine clinical care at Tata Medical Center (TMC).

Since the study involves only retrospective analysis of existing clinical records and archived pathological slides, there are no direct patient safety concerns associated with study participation. The digitization of archival H&E slides is a non-destructive process and does not alter the original specimens.

Patient confidentiality and data security are the primary safety considerations. All data will be anonymized prior to computational analysis, and access to identifiable data will be restricted to authorized study personnel only. The study will be conducted in accordance with the Declaration of Helsinki and applicable institutional and national ethical guidelines.

Any incidental findings of clinical significance identified during data review will be handled in accordance with institutional protocols.

6) Funding

This study is an investigator-initiated retrospective study conducted at Tata Medical Center (TMC), Kolkata. Funding support is currently being applied for through institutional and national research grant mechanisms. The study does not involve any commercial sponsorship or industry funding at this time. All computational resources required for AI model development will be sourced through institutional infrastructure and open-access platforms.

7) References

1. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. *Lancet*. 2014;384(9938):164–172.
2. Symmans WF, Wei C, Gould R, et al. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J Clin Oncol*. 2017;35(10):1049–1060.
3. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. *Lancet*. 2021;397(10286):1750–1769.
4. Dent R, Trudeau M, Pritchard KI, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin Cancer Res*. 2007;13(15):4429–4434.
5. Salgado R, Denkert C, Demaria S, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann Oncol*. 2015;26(2):259–271.
6. Denkert C, von Minckwitz G, Darb-Esfahani S, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol*. 2018;19(1):40–50.
7. Kather JN, Heij LR, Grabsch HI, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer*. 2020;1(8):789–799.
8. Chen RJ, Lu MY, Weng WH, et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. *Proc IEEE Int Conf Comput Vis*. 2021:4015–4025.
9. Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210.
10. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–1567.
11. Kleppe A, Skrede OJ, De Raedt S, et al. Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer*. 2021;21(3):199–211.
12. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal*. 2021;67:101813.
13. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25(8):1301–1309.
14. Lu MY, Williamson DFK, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570.
15. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med*. 2024;30(3):850–862.
16. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
17. Antonini M, Mattar A, Pereira TM, Oliveira LL, Teixeira MD, Amorim AG, Ferraro O, de Oliveira LC, Ramos MD, Cavalcante FP, Zerwes F. Pathologic complete response and breast cancer survival post-neoadjuvant chemotherapy: A systematic review and meta-analysis of real-world data. *Heliyon*. 2025 May 1;11(10).