

TITLE: Predicting Pathological Complete Response to NACT in Breast Cancer patients using High-Throughput Images and Machine Learning- retrospective

BACKGROUND:

Neoadjuvant chemotherapy (NACT) in breast cancer has significantly improved surgical feasibility and oncological outcomes, while enabling breast and axillary conservation in selected patients. Pathological complete response (pCR) is widely accepted as a surrogate marker for long-term outcomes, particularly in triple-negative breast cancer. However, reported pCR rates vary widely (10–60%) depending on molecular subtype, treatment regimen, and patient characteristics. Currently, the decision to offer NACT in India is primarily based on clinical stage and tumour biology, without reliable tools to predict individual response. Accurate prediction of pCR prior to initiating NACT could enable personalized treatment selection—identifying patients most likely to benefit from chemotherapy, sparing others from unnecessary toxicity, optimizing surgical planning, and reducing overall treatment costs. This study aims to develop an indigenous, affordable AI-based model tailored to the Indian population for predicting pCR following NACT in breast cancer.

STUDY DESIGN :

Retrospective analysis of prospectively collected data for model development and internal validation

INCLUSION CRITERIA:

- Age 18–70 years
- Operable breast cancer
- Core biopsy-proven TNBC
- Received NACT followed by definitive breast and axillary surgery

METHOD FOR MODEL DEVELOPMENT :

This study will develop and evaluate an AI model to predict pathological complete response (pCR; binary yes/no per a pre-specified, locked pathological definition) in triple-negative breast cancer using only pre-neoadjuvant chemotherapy (pre-NACT) data. The primary (Tier A) approach relies on baseline digitized H&E core-biopsy whole-slide images (WSIs) and routinely collected clinical covariates. From WSIs, in-house deep phenotyping pipelines generate interpretable morphology and cellular features (e.g., tumor-region localization, nuclear/cell-type composition, and mitotic activity surrogates) and aggregate patch-level representations to a slide-level pCR probability using structured, transformer-based models that incorporate spatial context. Model development will use patient-level separation with nested cross-validation for hyperparameter tuning and performance estimation, with appropriate methods for class imbalance and prespecified operating thresholds. Particular attention will be given to pCR label quality control and explicit TNBC eligibility criteria to mitigate noise arising from variability in TNBC/IHC definitions.

NUMBER OF SUBJECTS , STUDY SITE AND DURATION

Target sample size: 500 patients (70% for model development and 30% for internal validation), who have undergone treatment at TMC from period of 2011 to 2025

Data will be collected from departmental Redcap database and pathology slides will be retrieved from lab. Expected time period of study for model development and internal validation is 3 years.