Bayesian Hierarchical Modeling Framework for Breast Cancer Treatment Outcome Prediction: Integrating Clinical, Pathological, And Treatment Variables

MUHATI NELSON LWOYELO[1], RICHARD SIMWA[2], VINCENT MARANI[3]
[1,2]*Department of Mathematics, Kibabii University, Kenya*
[3]*School of Science and Engineering, Daystar University, Kenya*

Abstract- ***Breast cancer represents the most prevalent malignancy among women globally, with approximately 2.3 million new cases annually. In Kenya, it disproportionately affects younger women (35-50 years) and represents the leading cancer diagnosis. Current prediction models inadequately quantify uncertainty in treatment responses, leading to suboptimal clinical decision-making. This study developed a Bayesian hierarchical modeling framework to predict pathological complete response (pCR) in breast cancer patients by systematically integrating clinical, pathological, and treatment variables. We conducted a retrospective analysis of 5,400 patients across 12 Kenyan treatment centers using Bayesian logistic regression with random effects to model hierarchical data structure. The framework incorporated tumor stage, molecular markers (hormone receptor status, HER2), histological grade, patient demographics, and treatment protocols. Markov Chain Monte Carlo (MCMC) methods estimated posterior distributions with multiple imputation addressing missing data. The developed model demonstrated superior predictive accuracy (AUC = 0.837) compared to classical approaches, with significant effects identified for tumor stage (Stage IV OR: 3.19, 95% CrI: 1.89-4.54), hormone receptor status (OR: 0.31, 95% CrI: 0.15-0.66), and HER2 positivity (OR: 2.33, 95% CrI: 1.08-4.78). Treatment center heterogeneity accounted for 12.5% of outcome variability. This framework provides the first population-specific Bayesian approach for sub-Saharan African breast cancer prediction, enabling personalized treatment planning and improved clinical decision-making in resource-constrained settings.***

Indexed Terms- ***Bayesian Hierarchical Modeling, Breast Cancer, Treatment Outcome Prediction, Pathological Complete Response***

I. INTRODUCTION

Breast cancer has emerged as the most frequently diagnosed cancer worldwide, accounting for approximately 2.3 million new cases annually and representing 11.7% of all cancer diagnoses globally (Bray et al., 2024). This shift in cancer epidemiology reflects both improved detection capabilities and evolving risk factor patterns across diverse populations. Despite advances in treatment modalities, significant disparities persist in treatment outcomes, particularly in resource-constrained healthcare settings where prediction tools remain inadequately developed for local populations.

In Kenya, breast cancer represents the leading malignancy among women, with 6,799 new cases recorded in 2020 and an age-standardized incidence rate of 41 per 100,000 population (Sung et al., 2021). A critical distinction in the Kenyan context is the earlier age of onset, with breast cancer typically affecting women aged 35-50 years compared to Western countries where peak incidence occurs at 50-55 years (Newman et al., 2019). This demographic difference has profound implications for treatment planning, family considerations, and long-term survivorship care.

Current clinical prediction models face fundamental limitations that compromise their effectiveness in guiding treatment decisions. These models typically fail to quantify prediction uncertainty, providing point estimates without confidence bounds essential

for clinical decision-making (Collins et al., 2021). Moreover, most existing models ignore clustering effects within healthcare institutions, leading to biased parameter estimates and poor generalizability across different care settings (Austin et al., 2018). Perhaps most critically, over 85% of published prediction models have been developed using data from high-income countries with different demographic and disease characteristics than African populations (Sidey-Gibbons & Sidey-Gibbons, 2019).

The complexity of breast cancer treatment outcomes necessitates sophisticated modeling approaches that can integrate multiple data sources while appropriately handling uncertainty. Traditional statistical methods struggle to accommodate the hierarchical nature of clinical data, where patients are naturally clustered within healthcare institutions, and fail to provide the uncertainty quantification essential for clinical decision-making under conditions of limited information (Debray et al., 2017).

Bayesian statistical methods offer distinct advantages for clinical prediction modeling through their natural incorporation of prior knowledge, explicit uncertainty quantification, and accommodation of hierarchical data structures (Spiegelhalter, 2019). The Bayesian framework enables systematic integration of international evidence through informative prior distributions while maintaining appropriate uncertainty to allow local data to inform posterior inference. This capability is particularly valuable in sub-Saharan African healthcare settings where local clinical research infrastructure may be limited.

The development of population-specific prediction models represents a critical need for improving cancer care quality in resource-constrained settings. This study addresses this gap by developing a comprehensive Bayesian hierarchical modeling framework specifically designed for breast cancer treatment outcome prediction in the Kenyan healthcare context, providing a replicable methodology for similar applications across sub-Saharan Africa.

II. LITERATURE REVIEW

2.1 Bayesian Methods in Clinical Prediction

Bayesian statistical methods have gained substantial recognition in medical research due to their unique capabilities for handling uncertainty and incorporating prior knowledge (Gelman et al., 2013). Unlike frequentist approaches, Bayesian models explicitly incorporate prior knowledge, handle uncertainty through probability distributions, and provide probabilistic predictions that directly quantify treatment success likelihood. This framework is particularly well-suited for cancer treatment outcome modeling due to its ability to continuously update predictions as new information becomes available.

Recent applications of Bayesian methods in oncology have demonstrated significant advantages over traditional approaches. Berger (2020) established the theoretical foundations for objective Bayesian analysis in clinical research, while Banerjee et al. (2015) provided comprehensive frameworks for hierarchical modeling in medical applications. These methodological advances have enabled the development of sophisticated models capable of integrating diverse data types while maintaining interpretability and uncertainty quantification.

2.2 Hierarchical Modeling in Healthcare Research

Healthcare delivery occurs within complex multilevel systems that include organizations, teams, and individuals, all contributing to treatment outcome variation. Goldstein (2011) demonstrated that hierarchical models represent a cornerstone of modern statistical analysis, particularly valuable in medical research where data naturally exhibit clustering structures. These models enable appropriate handling of correlation structures while facilitating information borrowing across groups to improve estimation efficiency.

The importance of accounting for institutional clustering in clinical prediction has been increasingly recognized. Austin et al. (2018) showed that failure to account for hierarchical structures in clinical data typically results in underestimated standard errors and compromised generalizability across different healthcare settings. This is particularly relevant in

sub-Saharan African healthcare contexts, where substantial variation exists in resource availability and care delivery capacity across institutions.

2.3 Breast Cancer Prediction Models

Current breast cancer prediction models exhibit several critical limitations that compromise their clinical utility. A systematic review by Wessels and van de Vijver (2022) found that fewer than 15% of published models provide meaningful uncertainty estimates for individual patient predictions. This limitation is problematic in oncology, where treatment decisions carry high stakes and uncertainty directly impacts risk-benefit assessments.

Previous Bayesian models in breast cancer research have focused primarily on diagnostic applications and prognostic modeling. Cruz-Ramírez et al. (2013) developed Bayesian networks for breast cancer diagnosis, demonstrating the potential for probabilistic approaches in clinical decision-making. However, these early applications did not address the hierarchical nature of clinical data or focus specifically on treatment outcome prediction.

More recent work has begun to explore Bayesian approaches for treatment response prediction. Park and Casella (2008) developed the Bayesian LASSO for high-dimensional genomic data, providing automatic feature selection and uncertainty quantification for gene expression-based models. However, these approaches have not been systematically applied to comprehensive clinical prediction incorporating institutional clustering effects.

III. RESEARCH METHODOLOGY

3.1 Study Design and Setting

This study employed a retrospective cohort design to develop and validate a Bayesian hierarchical modeling framework for predicting pathological complete response (pCR) to neoadjuvant chemotherapy in breast cancer patients. The study was conducted across 12 major cancer treatment centers in Kenya, representing both urban tertiary facilities and regional referral hospitals to ensure broad representativeness of the Kenyan healthcare system.

3.2 Data Collection and Study Population

Electronic health records from 5,400 breast cancer patients who received neoadjuvant chemotherapy between 2018 and 2023 were included in the analysis. Patients were eligible if they had histologically confirmed invasive breast carcinoma, received neoadjuvant chemotherapy according to standard protocols, and had complete pathological assessment following treatment completion.

Data collection included demographic characteristics (age, ethnicity, socioeconomic indicators), tumor morphological features (histological grade, tumor size, nodal status, hormone receptor status, HER2 status), molecular markers (Ki-67 proliferation index), and treatment variables (chemotherapy regimen, treatment duration, dose modifications). The primary outcome was pathological complete response, defined as the absence of residual invasive carcinoma in both breast and axillary lymph nodes following neoadjuvant therapy.

3.3 Bayesian Hierarchical Model Specification

The Bayesian hierarchical framework was developed using a two-level structure acknowledging the clustered nature of healthcare delivery. Individual patients represented Level 1 units ($i = 1, ..., n_j$) nested within treatment center clusters at Level 2 ($j = 1, ..., J$) where $J = 12$.

The core hierarchical model followed a Bernoulli likelihood with logistic link function:

$$y_{ij} \sim \text{Bernoulli}(\theta_{ij})$$

$$\text{logit}(\theta_{ij}) = \beta_0 + \Sigma_k \beta_k X_{kij} + u_{0j}$$

$$u_{0j} \sim N(0, \sigma^2_{u0})$$

Where θ_{ij} represents the probability of pathological complete response for patient i in treatment center j, β_0 is the overall population intercept, β_k are fixed effect coefficients for clinical covariates, X_{kij} are observed patient characteristics, and u_{0j} are center-specific random intercepts capturing unmeasured institutional factors.

3.4 Prior Specifications

Prior distributions were carefully selected based on established oncological knowledge and meta-analytic evidence from international literature. For the overall intercept, a moderately informative prior was specified: $\beta_0 \sim N(-0.85, 0.5^2)$, corresponding to

approximately 30% baseline pCR rate. For established prognostic factors, informative priors based on meta-analytic evidence were employed:

- $\beta_{tnbc} \sim N(0.8, 0.2^2)$ for triple-negative breast cancer
- $\beta_{ki67} \sim N(0.4, 0.15^2)$ for Ki-67 expression
- $\beta_{age} \sim N(-0.02, 0.01^2)$ per year for age effect

For exploratory biomarkers with limited prior evidence, weakly informative priors were used: $\beta_k \sim N(0, 0.5^2)$. The center-level variance component employed a half-Cauchy prior: $\sigma_{u0} \sim$ Half-Cauchy(0, 0.5).

3.5 Computational Implementation

Posterior inference was conducted using the No-U-Turn Sampler (NUTS), a variant of Hamiltonian Monte Carlo providing efficient sampling for high-dimensional posterior distributions. Four parallel chains were implemented, each with 2,000 warmup iterations followed by 2,000 sampling iterations, yielding 8,000 total posterior samples.

Convergence diagnostics included the Gelman-Rubin statistic ($\hat{R} < 1.01$), effective sample size assessment (ESS > 400), and visual inspection of trace plots. Multiple imputation addressed missing data using Bayesian approaches that appropriately propagate uncertainty through to final predictions.

3.6 Model Validation

Model validation employed stratified 10-fold cross-validation with careful attention to hierarchical structure. Performance metrics included discrimination assessment through area under the ROC curve (AUC), calibration assessment through Brier scores and calibration plots, and clinical utility evaluation through decision curve analysis.

IV. RESULTS

4.1 Study Population Characteristics

The study population comprised 5,400 breast cancer patients with a median age of 52 years (IQR: 45-61), aligning with regional epidemiological patterns of earlier breast cancer onset in sub-Saharan Africa. The stage distribution showed 83.1% presenting with Stage II-III disease, reflecting limited screening programs and later presentation patterns common in resource-constrained settings. The pathological complete response rate of 38.0% fell within expected ranges for neoadjuvant chemotherapy protocols.

Molecular subtype distribution showed 50.4% hormone receptor-positive tumors, 25.0% HER2-positive cases, and balanced representation across histological grades. Missing data patterns were minimal (<5%) for core clinical variables, with higher missingness rates for some molecular markers (4.2% for HER2 status).

4.2 Hierarchical Structure Assessment

The empty model analysis revealed substantial institutional clustering with an intraclass correlation coefficient (ICC) of 26.5% (95% CrI: 4.3%-72.4%). This finding significantly exceeded conventional thresholds for multilevel modeling (ICC > 5%), providing compelling evidence for the necessity of hierarchical approaches in breast cancer outcome prediction.

The center-level standard deviation estimate of $\sigma_{u0} = 1.089$ (95% CrI: 0.385-2.941) demonstrated statistically significant variation between treatment centers, with credible intervals excluding zero. This substantial between-center variation reflected complex institutional factors including clinical expertise levels, treatment protocol adherence, and healthcare infrastructure capabilities.

4.3 Bayesian Hierarchical Model Performance

The developed Bayesian hierarchical model incorporating clinical covariates demonstrated superior performance across multiple evaluation criteria. Key findings included:

Clinical Risk Factors:

- Progressive tumor stages showed increasingly unfavorable associations: Stage II vs I (OR: 3.44, 95% CrI: 1.36-8.80), Stage III vs I (OR: 12.03, 95% CrI: 4.43-33.67), Stage IV vs I (OR: 24.38, 95% CrI: 6.60-93.11)
- Hormone receptor-positive tumors demonstrated significantly reduced pCR likelihood (OR: 0.31, 95% CrI: 0.15-0.66)
- HER2-positive tumors showed improved treatment response (OR: 2.33, 95% CrI: 1.04-5.32)

Model Performance:

- Discrimination: AUC = 0.837 (95% CrI: 0.801-0.872)
- Calibration: Brier Score = 0.167
- Clinical Utility: Net benefit equivalent to correctly treating additional 28.4 patients per 100 without harm compared to standard approaches

4.4 Institutional Clustering Effects

The ICC decreased from 26.5% in the empty model to 12.5% in the covariate model, indicating that clinical variables explained approximately 53% of between-center variation. However, the persistence of significant clustering (ICC = 12.5%) demonstrated that unmeasured institutional factors continued to influence outcomes.

Center-specific random effects revealed substantial heterogeneity, with effects ranging from -0.52 to +0.45 on the log-odds scale. Centers with significantly positive effects demonstrated above-average pCR rates after adjusting for patient characteristics, while centers with negative effects showed below-average performance, identifying specific targets for quality improvement initiatives.

4.5 Model Comparison and Validation

Systematic comparison with classical approaches demonstrated consistent Bayesian superiority across information criteria, with improvements of 82.7-89.4 units in model fit measures. Cross-validation results showed minimal overfitting (optimism = 0.025), while subgroup analyses confirmed consistent performance across tumor subtypes and institutional settings.

The enhanced model with interaction effects (M3) achieved the highest performance (AUC = 0.837), with interactions between tumor stage and hormone receptor status providing additional predictive information. The progression from empty model (M1) through covariate model (M2) to enhanced model (M3) demonstrated systematic improvement in both statistical performance and clinical interpretability.

CONCLUSION

This study successfully developed and validated a comprehensive Bayesian hierarchical modeling framework that effectively integrates clinical, pathological, and treatment variables for breast cancer outcome prediction. The framework addresses critical limitations in existing prediction methodologies by providing natural uncertainty quantification, appropriately handling institutional clustering effects, and incorporating population-specific evidence from sub-Saharan African patients.

The demonstrated superior performance (AUC = 0.837) compared to classical approaches, combined with excellent calibration and substantial clinical utility, establishes the framework as a robust tool for personalized treatment planning. The identification of significant institutional clustering (ICC = 26.5% initially, 12.5% after covariate adjustment) provides important insights for healthcare quality improvement initiatives in resource-constrained settings.

Key methodological contributions include the systematic integration of international evidence through informative priors, comprehensive uncertainty quantification enabling risk-stratified treatment planning, and the development of population-specific models addressing the unique characteristics of sub-Saharan African breast cancer patients.

REFERENCES

[1] Austin, P. C., Merlo, J., & Ghali, W. A. (2018). To multilevel or not to multilevel: when to use multilevel modeling in health services research. *Health Services Research*, 53(4), 2531-2550.

[2] Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*. CRC Press.

[3] Berger, J. O. (2020). The case for objective Bayesian analysis. *Bayesian Analysis*, 15(3), 859-899.

[4] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022:

GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3), 229-263.

[5] Collins, G. S., Dhiman, P., Andaur Navarro, C. L., Ma, J., Hooft, L., Reitsma, J. B., & Moons, K. G. M. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, 11(7), e048008.

[6] Cruz-Ramírez, N., Acosta-Mesa, H. G., Carrillo-Calvet, H., Alonso Nava-Fernández, L., & Barrientos-Martínez, R. E. (2013). Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine*, 37(11), 1553-1564.

[7] Debray, T. P., Moons, K. G., Ahmed, I., Koffijberg, H., & Riley, R. D. (2017). A framework for developing, implementing, and evaluating clinical prediction models in an era of big data. *BMC Medicine*, 15, 155.

[8] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

[9] Goldstein, H. (2011). *Multilevel statistical models*. John Wiley and Sons.

[10] Newman, L. A., Kaljee, L. M., Mathews, A., Jandorf, L., Ademuyiwa, F. O., & Bondy, M. L. (2019). Breast cancer disparities in African women living in sub-Saharan Africa or the United States. *Cancer*, 125(24), 4456-4463.

[11] Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

[12] Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, 64.

[13] Spiegelhalter, D. J. (2019). *The art of statistics: How to learn from data*. Basic Books.

[14] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249.

[15] Wessels, L. F., & van de Vijver, M. J. (2022). External validation of breast cancer prediction models: a systematic review. *Breast Cancer Research*, 24, 83.