COMPARING THE EFFECTIVENESS OF RISK FACTOR ABOUT HER-2 LOW WITH FLUORESCENCE IN SITU HYBRIDIZATION BREAST CANCER METASTASIS AMONG WOMAN AGED GREATER OR EQUAL TO 45 IN CHINA.

By: Yi Yang

Bachelor of Computer Science, Miami University, 2020

Thesis Advisor: Hai Shu, PhD

A Master’s thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Department of Biostatistics

School of Global Public Health, New York University

May 2024

1. **Abstract.**

1.1 Objective.

Breast cancer is a leading cause of mortality in women, and understanding its molecular characteristics, particularly Her-2 status, is crucial. This study focuses on Her-2 low breast cancer especially in elderly Chinese women. Using fluorescence in situ hybridization (FISH), the study identifies tumor size, number of lymph nodes, and CA125 as significant variables related to Her-2 low with FISH breast cancer. Machine learning methods enhance the understanding of disease progression and risk factors. The study sheds light on the implications of Her-2 low status for metastasis in this population, aiming to improve personalized interventions and treatment outcomes. These findings may enhance breast cancer care for elderly Chinese women with Her-2 low status.

1.2 Method.

This retrospective cohort study examines 1874 breast cancer patients from Beijing Pinggu Hospital, China, focusing on Her-2 low breast cancer with FISH. Outpatient and surveillance reviews were conducted post-treatment. The hypothesis is that no significant risk factor influences metastasis rates. BIC and Backward Selection were used for variable selection. Machine learning models (Random Forest, KNN, CART, XGBoost, Neural Networks) were employed, with ROC and AUC used for model evaluation. Importance analysis and MCC were used to assess variable importance. Results will enhance understanding of Her-2 low breast cancer with FISH and improve treatment outcomes for this population.

1.3 Result.

This study examines machine learning models for predicting breast cancer metastasis using a dataset of 676 samples. Variable selection via BIC and Backward Stepwise Selection methods led to Random Forest with two trees as the optimal model, achieving an AUC of 0.7048. Other models, including KNN, CART, XGBoost, and Neural Networks, also showed reasonable performance. CA125 emerged as the most influential predictor.

1.4 Discussion.

Variable selection methods identify important predictors, with Random Forest emerging as the most effective model (AUC=0.7048), followed by XGBoost (AUC=0.6786) and Logistic Regression (AUC=0.6939). CA125 is highlighted as a crucial predictor, along with Tumor\_size and Number\_of\_lymph\_nodes. These findings demonstrate the potential of machine learning in predicting metastasis and the importance of specific variables in the model.

1.5 Key Words.

Her-2 with Fish Breast Cancer, Metastasis, woman, age greater than 45, CA125, Machine Learning, Tumor Size, Number of lymph nodes.

1. **Introduction.**

2. 1 Background.

Breast cancer is the most common malignant tumor of women, and its mortality rate ranks second among female tumors even the comprehensive treatment level has been constantly improving.1 Breast tumors usually start from the ductal hyperproliferation, and then develop into benign tumors or even metastatic carcinomas after constantly stimulation by various carcinogenic factors.2

Her2 belongs to the Her/ErbB2/Neu family of transmembrane receptors, alongside Her1/EGFR, Her3, and Her4. These receptors are characterized by three distinct regions: an extracellular domain (ECD), a membrane-spanning region, and a cytoplasmic tyrosine kinase domain. 3 In 15-20% of breast cancers, the member of the HER receptor family, HER2, is overexpressed. 4 Although specific risk factors associated with economic development are largely unavoidable, the projected substantial rise in new cases of breast cancer underscores the urgency of incorporating this disease into future healthcare infrastructure planning at an early stage. 5 In the contemporary landscape, the increasing emphasis on Her-2 low expression breast cancer is propelled by the advancement of antibody-drug conjugates (ADC). Recent studies indicate that the Her-2 receptor is present to some extent on the cell membrane, drawing heightened attention to the exploration of its implications in breast cancer research. 6 Her-2 low is defined as Her-2 (1+) or Her-2 (2+) with ISH (-) according to the current clinical scoring system by most of the published data and ongoing clinical trials. 7–15 Several studies have reported poorer prognosis in Her-2 low positive breast cancer patients.16–19 But so far, no such results in Chinese patients are reported. In China, Her-2 low-positive breast cancers made up 54% of all patients, as calculated by a multi-center retrospective research. 5

Fluorescence in situ hybridization (FISH) serves as an alternative technique for evaluating HER2 status in clinical specimens. This method, directly quantifying the HER2 gene copy number, has proven to be a dependable approach for assessing HER2 status in formalin-fixed, paraffin-embedded clinical specimens. Its demonstrated high sensitivity and specificity rates underscore its reliability in determining the HER2 status, providing a valuable tool for accurate molecular characterization in clinical settings. 20

Presently, the identification of breast cancer metastasis hinges on several methods, including the assessment of clinical symptoms indicative of distant organ involvement, biopsies extracted from affected organs, radiological examinations, various imaging modalities, and the measurement of serum tumor markers.21,22 Furthermore, ASCO also advocates for the use of mammography as a means of early detection for relapse in breast cancer.23 The process of metastasis comprises a series of sequential steps. Failure to complete any of these steps will arrest the process. 24 Metastasis starts with the local invasion of surrounding host tissue by cells originating from the primary tumor and continues until the tumor cells invade and intravasate into blood or lymphatic vessels. 25,26 The tumor cells are disseminated via the bloodstream or the lymphatic vessels to distant organs. Consequently, the tumor cells undergo cell cycle arrest and adhere to capillary beds within the target organ, before extravasating into the organ parenchyma, proliferating and promoting angiogenesis within the organ.25 During the progression through these steps, the tumor cells must concurrently navigate the host's immune response and evade apoptotic signals to ensure their survival.25,27 Should the tumor cells successfully navigate through these steps, the process can be iterated, leading to the formation of secondary metastases, commonly referred to as 'metastasis of metastases.' 24,26

According to the recent study Dr.Seung from Samsung Medical Center in korea has shown that Tumor size, number of Lymph nodes,and CA125 are three significant variables that are related to Her-2 low with FISH breast cancer, and these three variables are also the exposure for this study.28 Many of the studies have mentioned that machine learning is an important methods for data analysis in public health and also to compare importance for the variables. The result for these data analysis can significantly Improve understanding of the disease with exposure.29,30

CA125 is an antigen associated with coelomic epithelium and serves as a valuable marker for monitoring residual disease in individuals undergoing chemotherapy for ovarian cancer. The interpretation of serum CA125 levels traditionally relies on a standard value of 35 U/ml. This benchmark was established by screening a diverse cohort of young blood donors from the general population, inclusive of women with intact reproductive systems. 31

Traditionally, the size of a breast cancer at the time of diagnosis has been regarded as a crucial factor influencing clinical outcomes. Nevertheless, the aggressive behavior exhibited by certain subtypes of breast cancer, even when they are small (≤1 cm in diameter), challenges the assumption that cancer size should be the sole consideration in treatment decisions. While an association between tumor size and lymph-node involvement holds for most tumor types, this correlation is not universally consistent. In specific subtypes of breast cancers, the relationship among tumor size, lymph-node status, and prognosis may indicate an underlying disproportionate connection between the number of cancer cells and their metastatic potential. This nuanced understanding urges a more comprehensive evaluation beyond tumor size alone when determining optimal treatment strategies for certain breast cancer subtypes. 32

Machine learning is vital in analyzing breast cancer metastasis, aiding in pattern recognition and outcome prediction from complex data. Metastasis, the cancer's spread to other body parts, significantly affects patient prognosis and treatment decisions. By analyzing factors like tumor characteristics and patient data, machine learning predicts metastasis likelihood, helping tailor treatments and improve outcomes. These algorithms also uncover key features driving metastasis, advancing our understanding and personalized treatment approaches.

2.2 Gaps and Significant.

Limited understanding of how Her-2 low breast cancer with FISH (fluorescence in situ hybridization) relates to breast cancer and metastasis in elderly Chinese women is the biggest gap. While extensive research has elucidated various aspects of breast cancer, the specific characteristics and implications of Her-2 low status in the context of FISH remain insufficiently explored, especially within the demographic of elderly Chinese women. This demographic subset often faces distinct challenges related to healthcare access, comorbidities, and treatment responses. By addressing this gap, the study aims to bridge the knowledge deficit and unravel critical insights into how Her-2 low breast cancer, as identified through FISH, contributes to the metastatic process in elderly Chinese women. Such findings would not only advance the understanding of breast cancer heterogeneity but also provide essential information for tailoring targeted and personalized interventions, thereby improving the prognosis and treatment outcomes for this specific population.

The most significant result of this study will highlight the crucial roles of risk factor for Her-2 low with FISH breast cancer metastasis in this particular population. Given the rising incidence of breast cancer in China and the increasing recognition of Her-2 low as a distinct entity, the outcomes of this study may have profound implications for patient management, contributing to improved clinical decision-making and ultimately enhancing the overall quality of breast cancer care for women in this demographic group.

1. **Method.**

3.1 Data Source and Analysis Tools.

The data for this study were sourced from icpsr.com, involving a retrospective cohort study conducted from May 14th, 2010, to October 26th, 2019, with a focus on 1874 breast cancer patients who underwent surgery at Beijing Pinggu Hospital in China. All patients received essential adjuvant or neoadjuvant treatments, such as chemotherapy, endocrine therapy, and radiotherapy, tailored to their pathology following the NCCN (National Comprehensive Cancer Network) breast cancer guidelines. Post-treatment, patients were scheduled for outpatient reviews every 3 months during the initial 2 years and every 6 months in the subsequent 3 years. After the completion of 5 years, surveillance reviews were conducted annually. All variables name, and corresponding values in the study are listed at the bottom of the proposal.33

In this study, R, Stata, and SAS programming languages will be used for data upload, data cleaning, and data analysis.

3.2 Research Question and Hypothesis.

The Research question for the study is How do different risk factors contribute to the metastasis of Her-2 low with FISH breast cancer, among women aged 45 or above in the Chinese population? So the outcome variable would be Metastasis, when the result is one, it means that the breast cancer has metastasized and vice versa. Exposure variables are Tumor Size, Number of Lymph nodes, CA125. In the exposure variables Tumor Size, Number of Lymph nodes, and CA125 are continuous variables.

The Hypothesis for the study is that there is no significant risk factor that can influence the metastasis rates of women who have Her-2 low with FISH breast cancer in the specified age group (≥45) in China.

3.3 Statistical Analysis Plan.

The study will employ Bayesian Information Criterion (BIC) and Backward Selection techniques for variable selection, aiming to identify the most precise variables that are suitable for machine learning models.

Bayesian Information Criterion (BIC) is a widely used model selection criterion that balances model fit and complexity by maximizing the likelihood function. It considers the number of parameters in the model and tends to favor simpler models to avoid overfitting. BIC offers several advantages, including its ability to handle model complexity and its consistency under certain conditions, ensuring that the probability of selecting the true model approaches 1 as the sample size approaches infinity. Additionally, BIC is computationally simple, making it easy to understand and implement. However, BIC tends to penalize complex models excessively, potentially overlooking some complex relationships in the data. It is also sensitive to sample size, often selecting overly simple models for small sample sizes, leading to underfitting. Furthermore, BIC's consistency results rely on assumptions such as the true model being among the candidate models and correctly specified, which may not hold in practice.

BIC=−2×LogLikelihood+k×log(n)

LogLikelihood is the maximized log-likelihood of the model given the data,

k is the number of parameters in the model, and

n is the number of observations in the data.

Backward selection is a feature selection method that iteratively fits the model and removes the least significant features to simplify the model. It helps reduce model complexity by eliminating features that have little impact on the response variable, thereby improving the model's interpretability and generalization ability. However, backward selection is a greedy algorithm and may get stuck in local optima, failing to find the global optimal solution. It can also be computationally expensive for large feature sets, as it requires repeatedly fitting the model and evaluating feature significance. Additionally, backward selection typically considers the significance of individual features and may overlook the impact of feature interactions on the model.

This study employed a range of machine learning algorithms, including Random Forest, k-Nearest Neighbors (KNN), Classification and Regression Trees (CART), XGBoost, and Neural Networks, for the purpose of analysis.

Random Forest (RF) is a powerful ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of individual trees. RF offers several advantages, including its ability to handle high-dimensional data and provide estimates of variable importance, which can aid in feature selection.34 Additionally, RF is less prone to overfitting compared to individual decision trees due to its ensemble nature. However, RF can be computationally expensive for large datasets, and its performance may degrade with highly imbalanced class distributions.

f(x) is the predicted output for input

is the number of trees in the forest

is the prediction of the i-th tree.

K-Nearest Neighbors (KNN) is a simple yet effective non-parametric classification and regression method. KNN makes predictions based on the majority class or average value of the k nearest data points in the feature space. One advantage of KNN is its simplicity and ease of implementation, making it suitable for various applications. However, KNN can be computationally demanding for extensive datasets, as it necessitates calculating distances between the target point and all other points in the dataset. Moreover, KNN is susceptible to irrelevant or redundant features, necessitating meticulous selection of the distance metric and the value of k.

is the predicted value for data point x.

is the target value of the i-th nearest neighbor of x.

Classification and Regression Trees (CART) are a popular decision tree algorithm used for both classification and regression tasks. CART recursively partitions the feature space into regions that minimize impurity, such as Gini impurity or entropy. One key advantage of CART is its interpretability, as the resulting tree can be visualized and easily understood. However, CART is prone to overfitting, especially with complex datasets, and can be unstable, leading to different trees with slight variations in the training data.

t is a node in the decision tree,

C is the number of classes,

p(i∣t) is the proportion of class i in node t.

XGBoost is an optimized implementation of gradient boosting machines, which sequentially combines weak learners (typically decision trees) to create a strong learner. XGBoost has gained popularity due to its efficiency, scalability, and high performance in various machine learning competitions. XGBoost implements regularization techniques to reduce overfitting and provides several hyperparameters for fine-tuning. However, XGBoost requires careful hyperparameter tuning and can be computationally expensive, especially for large datasets.

n is the number of training examples.

is the true label of the i-th training example.

is the predicted value for the i-th example.

K is the number of trees in the model.

represents the k-th tree.

is the regularization term for the k-th tree.

Neural Networks (NNs) are a class of models inspired by the structure and function of the human brain. NNs consist of interconnected nodes organized in layers, with each node applying a non-linear activation function to its inputs. NNs are capable of learning complex non-linear relationships in the data and are effective for large datasets with high-dimensional features. However, NNs require a large amount of data to train effectively and are computationally intensive. Additionally, NNs can be prone to overfitting if not regularized properly and can be challenging to interpret due to their complex architecture.

Following the application of machine learning techniques, this study will utilize Receiver Operating Characteristic (ROC) curves and the Area Under the ROC Curve (AUC) to determine the optimal models for both the product variables and the dataset.

The ROC curve is a graphical representation illustrating the performance of a binary classification model, showcasing the trade-off between sensitivity (true positive rate) and specificity (true negative rate) across various threshold values. This curve is constructed by plotting the true positive rate against the false positive rate at different threshold settings.

AUC, a scalar value ranging from 0 to 1, summarizes the overall performance of a model. A higher AUC value indicates better discrimination ability. This metric is widely used in evaluating binary classification models because it offers a comprehensive evaluation of the model's performance across all possible threshold settings, enabling comparison between different models and assessment of their predictive capabilities.

After choosing the best fit model, the study would use importance analysis and MCC to analyze the importance of three variables which are CA125, Tumor size, and Number of Lymph nodes.

Mean Decrease Accuracy and Mean Decrease Gini are metrics used in decision tree-based algorithms, like Random Forests, to assess the importance of variables/features in the model.

In Random Forests, this metric measures how much the accuracy of the model decreases when a particular variable is excluded. It's calculated by permuting the values of each predictor variable one at a time and then measuring the decrease in accuracy. A higher decrease indicates that the variable is more important for the model's accuracy.

Gini impurity is a measure of how often a randomly chosen element would be incorrectly classified. Mean Decrease Gini is computed by calculating the total decrease in node impurity (weighted by the probability of reaching that node) due to splits over a given predictor, averaged over all trees in the ensemble. It's a measure of how much each variable contributes to the homogeneity of nodes and, consequently, to the overall quality of the splits in the tree. Higher values indicate more important variables.

MCC stands for Matthews Correlation Coefficient, which is a measure used to assess the quality of binary (two-class) classification models, particularly when the classes are imbalanced. It takes into account true positives, true negatives, false positives, and false negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes.

The MCC is calculated as:

TP is the number of true positives

TN is the number of true negatives

FP is the number of false positives

FN is the number of false negatives.

The MCC ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and observation. In general, higher MCC values indicate better performance of the classification model.

1. **Results.**

4.1 Data Preparation.

The dataset consisted of 676 samples, with 331 in the training set, 203 in the testing set, and 142 in the validation set.

The Univariable analysis table (table 1) presents demographic and clinical characteristics of 676 cancer patients, including age, biomarker levels (CA125, CA15-3), interval between diagnosis and treatment, and various cancer-related factors. The mean age of patients was 57.0 years, with a range of 22.0 to 90.0 years. CA125 levels were significantly higher in metastatic cases (mean 86.2 U/ml) compared to non-metastatic cases (mean 12.3 U/ml). Similarly, CA15-3 levels were higher in metastatic cases (mean 97.6 U/ml) than in non-metastatic cases (mean 10.1 U/ml). Other variables, such as tumor size and number of lymph nodes, also showed differences between metastatic and non-metastatic cases. These findings provide valuable insights into the characteristics of the patient population and can help in predicting metastasis in cancer patients.

4.2 Variable Selection.

BIC Model: The best model selected by the Bayesian Information Criterion (BIC) was metastasis ~ CA125 + interval + LN\_2 + Cancer\_type\_9 + Number\_of\_lymph\_nodes + Tumor\_size.

The best model selected by the Backward Stepwise Selection method was metastasis ~ Number\_of\_lymph\_nodes + CA125 + interval + survey\_type\_3 + survey\_type\_4 + Cancer\_type\_3 + Cancer\_type\_4 + Molecular\_typing\_1 + Molecular\_typing\_2 + LN\_1 + PR\_1 + ER\_1 + Cancer\_type\_9 + Molecular\_typing\_4 + Tumor\_size.

4.3 Model Evaluation.

Through cross-validation, Random Forest achieved its optimal performance with two trees, yielding an AUC of 0.7048. K-Nearest Neighbors (KNN) also attained its highest performance with K set to two, resulting in an AUC of 0.6403. In contrast, Classification and Regression Trees (CART) achieved an AUC of 0.5399, showcasing comparatively lower predictive accuracy. XGBoost, when configured with 50 nrounds, yielded the highest AUC of 0.6786. Neural Networks, on the other hand, achieved an AUC of 0.5604. When considering the AUC values collectively, Random Forest with two trees emerges as the most suitable model for this study, suggesting its efficacy for further variable and model analysis in classification tasks.The comparison of all AUC graphs is shown in Figure 1.

4.4 Performance Metrics(MCC).

The model correctly classified 94.09% of all instances (both metastasis and non-metastasis cases) in the dataset.

the proportion of correctly predicted metastasis cases (true positives) among all instances predicted as metastasis.

Recall (also known as sensitivity) represents the proportion of correctly predicted metastasis cases (true positives) among all actual metastasis cases. The model correctly identified 42.86% of all actual metastasis cases.

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is useful when the dataset is imbalanced, as it considers both false positives and false negatives. The F1 score of 0.3333 indicates a moderate balance between precision and recall.

4.5 Variable Importance (Random Forest).

The presented results depict variable importance measures within a predictive model, with %IncMSE denoting the rise in mean squared error and IncNodePurity signifying the increase in node purity upon the inclusion of a specific variable. Notably, CA125 exhibits a substantial impact, with a %IncMSE of 6.62 and an IncNodePurity of 17.29, indicating its strong predictive power. Conversely, Tumor\_size demonstrates a negligible impact on mean squared error (-0.09) but a moderate increase in node purity (5.77). Similarly, the Number\_of\_lymph\_nodes variable shows a slightly negative effect on mean squared error (-0.21) but a meaningful improvement in node purity (4.17). These findings contribute valuable insights into the relative importance of each variable in the predictive performance of the model.

1. **Discussion.**

The analysis conducted in this study aimed to predict metastasis in cancer patients using various machine learning models. The dataset was divided into training, testing, and validation sets to develop and evaluate the models.

The BIC model and Backward Stepwise Selection model identified different sets of variables as important predictors for metastasis. The BIC model selected CA125, interval, LN\_2, and Cancer\_type\_9, while the Backward Stepwise Selection model chose CA125, interval, Number\_of\_lymph\_nodes, and several others. This discrepancy suggests that different variable selection methods may lead to different sets of important predictors.

Among the models evaluated, Random Forest emerged as the most effective, achieving an AUC of 0.7048. This indicates that Random Forest is a promising model for predicting metastasis in this study. XGBoost also showed good predictive performance with an AUC of 0.6786, slightly lower than Random Forest. Logistic Regression, particularly the BIC model, performed reasonably well with an AUC of 0.6939.

However, models such as K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), and Neural Network showed lower performance, suggesting that they may not be as suitable for predicting metastasis in this context.

The variable importance analysis reveals compelling insights into the predictive capabilities of CA125, Tumor\_size, and Number\_of\_lymph\_nodes in our model. CA125 emerges as a crucial predictor, displaying a substantial impact on both the reduction in mean squared error (%IncMSE of 6.62) and the improvement in node purity (IncNodePurity of 17.29). This underscores its pivotal role in enhancing the model's predictive accuracy and reliability. Despite Tumor\_size exhibiting a slight negative impact on %IncMSE (-0.09), it contributes significantly to node purity (5.77), indicating its importance as a predictor. Similarly, Number\_of\_lymph\_nodes shows a modest negative impact on %IncMSE (-0.21) but contributes positively to node purity (4.17). These findings emphasize the critical role of CA125, and to a lesser extent, Tumor\_size and Number\_of\_lymph\_nodes, in our predictive model, suggesting their value as key variables in predicting the outcome.

The primary bias in this study is rooted in the absence of comprehensive demographic data, which poses a significant limitation to the overall validity and reliability of the findings. Without a thorough understanding of the diverse characteristics of the study population, the results may not accurately represent the broader demographic landscape. This lack of inclusivity not only hampers the generalizability of the study but also raises ethical concerns, as it may inadvertently perpetuate disparities or overlook important nuances within specific subgroups. Moreover, the narrow demographic scope further compounds these issues, limiting the study's applicability to a more diverse or representative population. Consequently, addressing these biases becomes imperative to ensure the study's ethical soundness and enhance the external validity of its conclusions.

A significant methodological limitation in this study revolves around the dependence on the quality and quantity of the data used. The reliability and validity of the study's results are intricately tied to the accuracy and completeness of the data collected, introducing a potential source of bias and uncertainty. Moreover, the study's ability to generalize findings to new, unseen data remains uncertain, as the performance on unfamiliar datasets is not guaranteed. This limitation underscores the importance of cautious interpretation and generalization of the study's outcomes, emphasizing the need for robust validation measures and consideration of potential biases stemming from data-related challenges.

Reference

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71:209–49.
2. Sun Y-S, Zhao Z, Yang Z-N, et al. Risk factors and preventions of breast cancer. *International Journal of Biological Sciences*. 2017;13(11):1387-1397. doi:10.7150/ijbs.21635
3. Windy Dean-Colomb, Francisco J. Esteva. Her2-positive breast cancer: Herceptin and beyond, European Journal of Cancer. Volume 44. Issue 18. 2008. Pages 2806-2812. ISSN 0959-8049,<https://doi.org/10.1016/j.ejca.2008.09.013>.
4. Sapino A, Goia M, Recupero D, Marchiò C. Current Challenges for HER2 Testing in Diagnostic Pathology: State of the Art and Controversial Issues. Front Oncol. 2013;3:129. Published 2013 May 21. doi:10.3389/fonc.2013.00129.
5. Shui R, Liang X, Li X, et al. Hormone Receptor and Human Epidermal Growth Factor Receptor 2 Detection in Invasive Breast Carcinoma: A Retrospective Study of 12,467 Patients From 19 Chinese Representative Clinical Centers. Clin Breast Cancer. 2020;20(1):e65-e74. doi:10.1016/j.clbc.2019.07.013
6. Onsum MD, Geretti E, Paragas V, et al. Single-cell quantitative HER2 measurement identifies heterogeneity and distinct subgroups within traditionally defined HER2-positive patients. *The American Journal of Pathology*. 2013;183(5):1446-1460. doi:10.1016/j.ajpath.2013.07.015
7. Banerji U, van Herpen CM, Saura C, et al. Trastuzumab duocarmazine in locally advanced and metastatic solid tumours and HER2-expressing breast cancer: A phase 1 dose-escalation and dose-expansion study. *The Lancet Oncology*. 2019;20(8):1124-1135. doi:10.1016/s1470-2045(19)30328-6
8. Chick RC, Clifton GT, Hale DF, et al. Subgroup analysis of Nelipepimut-S Plus GM-CSF combined with trastuzumab versus trastuzumab alone to prevent recurrences in patients with high-risk, HER2 low-expressing breast cancer. *Clinical Immunology*. 2021;225:108679. doi:10.1016/j.clim.2021.108679
9. Mittendorf EA, Lu B, Melisko M, et al. Efficacy and safety analysis of Nelipepimut-s vaccine to prevent breast cancer recurrence: A randomized, multicenter, phase III clinical trial. *Clinical Cancer Research*. 2019;25(14):4248-4254. doi:10.1158/1078-0432.ccr-18-2867
10. Modi S, Saura C, Yamashita T, et al. Trastuzumab Deruxtecan in previously treated her2-positive breast cancer. *New England Journal of Medicine*. 2020;382(7):610-621. doi:10.1056/nejmoa1914510
11. Schneeweiss A, Park-Simon T-W, Albanell J, et al. Phase IB study evaluating safety and clinical activity of the anti-HER3 antibody lumretuzumab combined with the anti-HER2 antibody pertuzumab and paclitaxel in her3-positive, HER2-low metastatic breast cancer. *Investigational New Drugs*. 2018;36(5):848-859. doi:10.1007/s10637-018-0562-4
12. Clifton GT, Peace KM, Holmes JP, et al. Initial safety analysis of a randomized phase II trial of nelipepimut-S + GM-CSF and trastuzumab compared to trastuzumab alone to prevent recurrence in breast cancer patients with HER2 low-expressing tumors. *Clinical Immunology*. 2019;201:48-54. doi:10.1016/j.clim.2019.02.011
13. Pondé N, Aftimos P, Piccart M. Antibody-Drug Conjugates in Breast Cancer: a Comprehensive Review. *Curr Treat Options Oncol*. 2019;20(5):37. Published 2019 Apr 1. doi:10.1007/s11864-019-0633-6
14. Rinnerthaler G, Gampenrieder SP, Greil R. HER2 Directed Antibody-Drug-Conjugates beyond T-DM1 in Breast Cancer. *Int J Mol Sci*. 2019;20(5):1115. Published 2019 Mar 5. doi:10.3390/ijms20051115
15. Modi S, Jacot W, Yamashita T, et al. Trastuzumab Deruxtecan in previously treated HER2-low advanced breast cancer. *New England Journal of Medicine*. 2022;387(1):9-20. doi:10.1056/nejmoa2203690
16. Camp RL, Dolled-Filhart M, King BL, Rimm DL. Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. *Cancer Res*. 2003;63(7):1445-1448.
17. Gilcrease MZ, Woodward WA, Nicolas MM, et al. Even low-level HER2 expression may be associated with worse outcome in node-positive breast cancer. *Am J Surg Pathol*. 2009;33(5):759-767. doi:10.1097/PAS.0b013e31819437f9
18. Eggemann H, Ignatov T, Burger E, et al. Moderate HER2 expression as a prognostic factor in hormone receptor positive breast cancer. *Endocr Relat Cancer*. 2015;22(5):725-733. doi:10.1530/ERC-15-0335
19. Rossi V, Sarotto I, Maggiorotto F, et al. Moderate immunohistochemical expression of HER-2 (2+) without HER-2 gene amplification is a negative prognostic factor in early breast cancer. *Oncologist*. 2012;17(11):1418-1425. doi:10.1634/theoncologist.2012-0194
20. Pauletti G, Godolphin W, Press MF, Slamon DJ. Detection and quantitation of HER-2/neu gene amplification in human breast cancer archival material using fluorescence in situ hybridization. *Oncogene*. 1996;13(1):63-72.
21. Lacroix M. Significance, detection and markers of disseminated breast cancer cells. *Endocrine-Related Cancer*. 2006;13(4):1033-1067. doi:10.1677/erc-06-0001
22. Sun Y-F, Yang X-R, Zhou J, Qiu S-J, Fan J, Xu Y. Circulating tumor cells: Advances in detection methods, biological issues, and clinical relevance. *Journal of Cancer Research and Clinical Oncology*. 2011;137(8):1151-1173. doi:10.1007/s00432-011-0988-y
23. Khatcheressian JL, Wolff AC, Smith TJ, et al. American Society of Clinical Oncology 2006 update of the breast cancer follow-up and management guidelines in the adjuvant setting. *Journal of Clinical Oncology*. 2006;24(31):5091-5097. doi:10.1200/jco.2006.08.8575
24. Poste G, Fidler IJ. The pathogenesis of cancer metastasis. *Nature*. 1980;283(5743):139-146. doi:10.1038/283139a0
25. Hunter KW, Crawford NP, Alsarraj J. Mechanisms of metastasis. *Breast Cancer Research*. 2008;10(S1). doi:10.1186/bcr1988
26. Talmadge JE, Fidler IJ. AACR Centennial Series: The Biology of Cancer Metastasis: Historical Perspective. *Cancer Research*. 2010;70(14):5649-5669. doi:10.1158/0008-5472.can-10-1040
27. Fidler IJ, Gersten DM, Hart IR. The biology of cancer invasion and metastasis. *Adv Cancer Res*. 1978;28:149-250. doi:10.1016/s0065-230x(08)60648-x
28. Min SK, Lee SK, Woo J, et al. Relation Between Tumor Size and Lymph Node Metastasis According to Subtypes of Breast Cancer. *J Breast Cancer*. 2021;24(1):75-84. doi:10.4048/jbc.2021.24.e4
29. Fountain FF, Tolley E, Chrisman CR, Self TH. Isoniazid hepatotoxicity associated with treatment of latent tuberculosis infection. *Chest*. 2005;128(1):116-123. doi:10.1378/chest.128.1.116
30. Ong SK, Haruyama R, Yip CH, et al. Feasibility of Monitoring Global Breast Cancer Initiative Framework Key Performance Indicators in 21 Asian National Cancer Centers Alliance member countries. *eClinicalMedicine*. 2024;67:102365. doi:10.1016/j.eclinm.2023.102365
31. Alagoz T, Buller RE, Berman M, Anderson B, Manetta A, DiSaia P. What is a normal CA125 level? *Gynecologic Oncology*. 1994;53(1):93-97. doi:10.1006/gyno.1994.1093
32. Foulkes WD, Reis-Filho JS, Narod SA. Tumor size and survival in breast cancer—a reappraisal. *Nature Reviews Clinical Oncology*. 2010;7(6):348-353. doi:10.1038/nrclinonc.2010.39
33. Li M, Li X, Xu F, Jiang H. *A retrospective 10 year survival analysis of 1340 HER-2 low positive breast cancer patients*. Published online 2023. doi:10.21203/rs.3.rs-3028676/v1
34. Althiban AS, Alharbi HM, Al Khuzayem LA, Eassa FE. Predicting Software Defects in Hybrid MPI and OpenMP Parallel Programs Using Machine Learning. Electronics. 2024; 13(1):182. https://doi.org/10.3390/electronics13010182

Table 1. Univariable Analysis table for all variables with Metastasis.

|  | **0**  **(N=609)** | **1**  **(N=67)** | **Overall**  **(N=676)** |
| --- | --- | --- | --- |

| **Age** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 57.1 (12.7) | 56.4 (14.1) | 57.0 (12.9) |
| Median [Min, Max] | 57.0 [22.0, 90.0] | 57.0 [30.0, 84.0] | 57.0 [22.0, 90.0] |

| **CA125** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 12.3 (12.3) | 86.2 (228) | 19.6 (75.5) |
| Median [Min, Max] | 9.92 [2.30, 196] | 15.4 [3.60, 1560] | 10.1 [2.30, 1560] |

| **CA15-3** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 10.1 (8.03) | 97.6 (492) | 18.8 (156) |
| Median [Min, Max] | 8.30 [0.300, 101] | 13.4 [2.00, 4000] | 8.48 [0.300, 4000] |

| **interval** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 38.9 (29.0) | 40.3 (28.9) | 39.1 (29.0) |
| Median [Min, Max] | 35.0 [0, 123] | 40.0 [0, 113] | 35.0 [0, 123] |

| **new\_assistance** |  |  |  |
| --- | --- | --- | --- |
| 0 | 555 (91.1%) | 56 (83.6%) | 611 (90.4%) |
| 1 | 54 (8.9%) | 11 (16.4%) | 65 (9.6%) |

| **survey\_type** |  |  |  |
| --- | --- | --- | --- |
| 1 | 6 (1.0%) | 2 (3.0%) | 8 (1.2%) |
| 3 | 38 (6.2%) | 3 (4.5%) | 41 (6.1%) |
| 4 | 541 (88.8%) | 61 (91.0%) | 602 (89.1%) |
| 5 | 14 (2.3%) | 1 (1.5%) | 15 (2.2%) |
| 7 | 10 (1.6%) | 0 (0%) | 10 (1.5%) |

| **Cancer\_type** |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 (0.3%) | 3 (4.5%) | 5 (0.7%) |
| 10 | 44 (7.2%) | 3 (4.5%) | 47 (7.0%) |
| 11 | 9 (1.5%) | 1 (1.5%) | 10 (1.5%) |
| 12 | 31 (5.1%) | 2 (3.0%) | 33 (4.9%) |
| 2 | 457 (75.0%) | 53 (79.1%) | 510 (75.4%) |
| 3 | 11 (1.8%) | 1 (1.5%) | 12 (1.8%) |
| 4 | 8 (1.3%) | 0 (0%) | 8 (1.2%) |
| 5 | 15 (2.5%) | 1 (1.5%) | 16 (2.4%) |
| 6 | 10 (1.6%) | 0 (0%) | 10 (1.5%) |
| 7 | 8 (1.3%) | 1 (1.5%) | 9 (1.3%) |
| 8 | 12 (2.0%) | 1 (1.5%) | 13 (1.9%) |
| 9 | 2 (0.3%) | 1 (1.5%) | 3 (0.4%) |

| **Neural\_invasion** |  |  |  |
| --- | --- | --- | --- |
| 0 | 524 (86.0%) | 52 (77.6%) | 576 (85.2%) |
| 1 | 85 (14.0%) | 15 (22.4%) | 100 (14.8%) |

| **Lymphatic\_or\_blood\_vascular\_tumor\_emboli** |  |  |  |
| --- | --- | --- | --- |
| 0 | 401 (65.8%) | 35 (52.2%) | 436 (64.5%) |
| 1 | 208 (34.2%) | 32 (47.8%) | 240 (35.5%) |

| **Size\_greater\_than\_2** |  |  |  |
| --- | --- | --- | --- |
| 0 | 333 (54.7%) | 25 (37.3%) | 358 (53.0%) |
| 1 | 276 (45.3%) | 42 (62.7%) | 318 (47.0%) |

| **T\_stage** |  |  |  |
| --- | --- | --- | --- |
| 1 | 335 (55.0%) | 3 (4.5%) | 338 (50.0%) |
| 2 | 245 (40.2%) | 4 (6.0%) | 249 (36.8%) |
| 3 | 20 (3.3%) | 3 (4.5%) | 23 (3.4%) |
| 4 | 9 (1.5%) | 57 (85.1%) | 66 (9.8%) |

| **M\_stage** |  |  |  |
| --- | --- | --- | --- |
| 0 | 350 (57.5%) | 25 (37.3%) | 375 (55.5%) |
| 1 | 163 (26.8%) | 17 (25.4%) | 180 (26.6%) |
| 2 | 58 (9.5%) | 12 (17.9%) | 70 (10.4%) |
| 3 | 38 (6.2%) | 13 (19.4%) | 51 (7.5%) |

| **TNM\_stage** |  |  |  |
| --- | --- | --- | --- |
| 2 | 495 (81.3%) | 3 (4.5%) | 498 (73.7%) |
| 3 | 105 (17.2%) | 7 (10.4%) | 112 (16.6%) |
| 4 | 9 (1.5%) | 57 (85.1%) | 66 (9.8%) |

| **Tumor\_size** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 2.34 (1.23) | 3.05 (2.06) | 2.41 (1.35) |
| Median [Min, Max] | 2.00 [0.400, 10.0] | 2.50 [0.800, 15.5] | 2.00 [0.400, 15.5] |

| **Number\_of\_lymph\_nodes** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 2.29 (5.99) | 4.87 (7.47) | 2.54 (6.19) |
| Median [Min, Max] | 0 [0, 60.0] | 2.00 [0, 38.0] | 0 [0, 60.0] |

| **LN** |  |  |  |
| --- | --- | --- | --- |
| 1 | 49 (8.0%) | 1 (1.5%) | 50 (7.4%) |
| 2 | 559 (91.8%) | 66 (98.5%) | 625 (92.5%) |
| 3 | 1 (0.2%) | 0 (0%) | 1 (0.1%) |

| **ER** |  |  |  |
| --- | --- | --- | --- |
| 1 | 541 (88.8%) | 57 (85.1%) | 598 (88.5%) |
| 2 | 68 (11.2%) | 10 (14.9%) | 78 (11.5%) |

| **PR** |  |  |  |
| --- | --- | --- | --- |
| 1 | 516 (84.7%) | 48 (71.6%) | 564 (83.4%) |
| 2 | 93 (15.3%) | 19 (28.4%) | 112 (16.6%) |

| **Ki67** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 0.257 (0.190) | 0.315 (0.201) | 0.263 (0.192) |
| Median [Min, Max] | 0.200 [0, 0.900] | 0.300 [0.0300, 0.900] | 0.200 [0, 0.900] |

| **Molecular\_typing** |  |  |  |
| --- | --- | --- | --- |
| 1 | 223 (36.6%) | 15 (22.4%) | 238 (35.2%) |
| 2 | 321 (52.7%) | 43 (64.2%) | 364 (53.8%) |
| 3 | 2 (0.3%) | 0 (0%) | 2 (0.3%) |
| 4 | 63 (10.3%) | 9 (13.4%) | 72 (10.7%) |

| **KI67** |  |  |  |
| --- | --- | --- | --- |
| Mean (SD) | 25.7 (19.0) | 31.5 (20.1) | 26.3 (19.2) |
| Median [Min, Max] | 20.0 [0, 90.0] | 30.0 [3.00, 90.0] | 20.0 [0, 90.0] |

Figure 1. Comparison of AUC of All Models

