



Unsupervised Machine Learning for Five-Year Survival Prediction in Non-metastatic Breast Cancer Patients Treated With Chemotherapy, Radiotherapy, and Surgery

Farzaneh Allaveisi¹, Javid Abdolmohammadi^{2,3}, Hanieh Parvaz⁴, Hamid Ghaznavi^{5,*}

¹ Department of Radiotherapy and Nuclear Medicine, Faculty of Paramedicine, Kurdistan University of Medical Sciences, Sanandaj, Iran

² Department of Medical Physics and Biomedical Engineering, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran

³ Department of Radiology, School of Allied Sciences, Kurdistan University of Medical Sciences, Sanandaj, Iran

⁴ Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

⁵ Social and Biological Network Analysis Laboratory, Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

*Corresponding Author: Social and Biological Network Analysis Laboratory, Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran. Email: ghaznavi2691@gmail.com

Received: 17 December, 2025; Revised: 25 February, 2026; Accepted: 13 April, 2026

Abstract

Background: Breast cancer patients often receive a combination of surgery, chemotherapy, and radiotherapy; however, predicting post-treatment survival remains challenging. Although machine learning methods offer new opportunities for risk stratification, supervised models are constrained by incomplete data and predefined targets. By contrast, unsupervised approaches, such as association rule mining (ARM), can uncover hidden and clinically meaningful patterns in complex breast cancer datasets.

Objectives: This study used ARM to identify patterns linking patient and tumor characteristics, as well as treatment details, to five-year survival or mortality in patients with non-metastatic breast cancer.

Methods: We retrospectively analyzed clinical data from 198 patients with breast cancer. Continuous variables, including age, tumor size, and dose, were binned, and all variables were converted into transaction-style records. Apriori-based ARM was applied with thresholds of support $\geq 5\%$, confidence $\geq 60\%$, and lift > 1 to extract the top ten rules predicting survival status as "Dead" or "Alive."

Results: Association rule mining identified two high-risk profiles for five-year mortality: 1) HER2-positive invasive ductal carcinoma treated with standard-dose radiotherapy and/or chemotherapy and 2) medium-sized tumors (2 - 5 cm) with grade 2 pathology treated with mastectomy and chemotherapy. Two low-risk profiles with excellent five-year survival were also identified: 1) patients aged 50 - 69 years with small (< 2 cm), PR-positive tumors and negative margins and 2) patients aged 50 - 69 years with grade 3, HER2-negative, PR-positive tumors treated with mastectomy or standard-dose radiotherapy. Correlation analyses confirmed tumor size and chemotherapy as the strongest predictors of survival for ARM.

Conclusions: Association rule mining identified distinct combinations of clinical and treatment factors that differentiated high-risk from low-risk breast cancer patients receiving combined therapies. These findings may help clinicians tailor follow-up intensity and treatment planning. Future studies should prospectively validate these rules and consider incorporating genomic or imaging data to further refine predictions.

Keywords: Breast Cancer, Survival Prediction, Association Rule Mining, Unsupervised Machine Learning

1. Background

Breast cancer (BC) remains the most commonly diagnosed cancer and the leading cause of cancer-

related death among women worldwide (1). One of the key challenges in BC management is its marked heterogeneity across genomic, epigenomic, transcriptomic, and proteomic levels (2, 3). This biological complexity results in substantial variation in

Copyright © 2026, Allaveisi et al. This open-access article is available under the Creative Commons Attribution 4.0 (CC BY 4.0) International License (<https://creativecommons.org/licenses/by/4.0/>), which allows for unrestricted use, distribution, and reproduction in any medium, provided that the original work is properly cited.

How to Cite: Allaveisi F, Abdolmohammadi J, Parvaz H, Ghaznavi H. Unsupervised Machine Learning for Five-Year Survival Prediction in Non-metastatic Breast Cancer Patients Treated With Chemotherapy, Radiotherapy, and Surgery. Int J Cancer Manag. 2026;19(1):e169145. doi: <https://doi.org/10.5812/ijcm-169145>

tumor growth, treatment response, and patient outcomes (4). Prognosis is influenced by several factors, including patient characteristics such as age; tumor features such as size and lymph node involvement (5, 6); and molecular biomarkers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status (7). However, these traditional markers often provide an incomplete assessment, and their accuracy in predicting survival or treatment response remains limited.

Recent advances in breast cancer treatment, including chemotherapy, radiotherapy, and surgery, have significantly improved survival and quality of life. Breast-conserving surgery (BCS) followed by adjuvant radiotherapy achieves locoregional recurrence rates of 2% - 3% for ER-positive and HER2-positive subtypes and 5% for triple-negative breast cancer (TNBC), comparable to outcomes after mastectomy alone (8, 9). Chemotherapy, particularly for TNBC and HER2-positive subtypes, not only improves pathologic complete response rates but can also convert some mastectomy candidates to BCS candidates (10, 11). Standard chemotherapy regimens often combine taxanes with anthracyclines, whereas platinum agents may offer additional benefit in TNBC, although their long-term impact remains under investigation (12-15). Radiotherapy remains essential after BCS, with hypofractionated whole-breast regimens preferred for early-stage disease and partial or accelerated protocols increasingly used for low-risk patients because of their favorable toxicity profiles (16-18). Postmastectomy radiotherapy is widely recommended for patients with 4 or more positive lymph nodes, and evidence also supports regional nodal irradiation in node-positive cases; however, the value of internal mammary node irradiation remains debated because of its limited benefit and potential for added toxicity (19-21).

Despite these therapeutic advances, predicting which patients will benefit most from specific treatments or which may relapse remains a major challenge. Tumor heterogeneity continues to limit the predictive value of traditional biomarkers such as ER, PR, and HER2, and newer biomarkers, such as circulating tumor DNA and immune gene signatures, remain limited by insufficient validation for widespread clinical use (22). Integrating multi-omics data, including genomics, proteomics, transcriptomics, and imaging, could improve risk stratification; however, this approach is challenging because of missing data, small sample sizes, and inconsistencies across datasets, all of which limit machine learning model performance (23). Predicting response to radiotherapy or chemotherapy is

particularly complex because validated radiosensitivity signatures are lacking and chemoresistance mechanisms are highly variable (24, 25). Thus, although outcomes have improved, the need for precise, personalized prediction tools in BC care remains unmet.

Recent advances in artificial intelligence-based breast cancer screening have demonstrated the ability to automatically detect suspicious features and uncover patterns in imaging data, achieving diagnostic performance comparable to, or even exceeding, that of radiologists (26-28). Among these methods, unsupervised learning approaches such as ARM have emerged as powerful tools for identifying relationships that are not explicitly labeled or predefined. For example, temporal ARM has been used to identify demographic and treatment patterns associated with survival in SEER-Medicare datasets (29). These studies indicate that unsupervised learning approaches such as ARM can provide clinically meaningful insights from large, heterogeneous datasets that might be missed by traditional statistical models.

2. Objectives

Building on this foundation, the present study applied ARM as an unsupervised machine learning approach to investigate survival outcomes in patients with breast cancer treated with combined modalities, including chemotherapy, radiotherapy, and surgery. By analyzing co-occurrence patterns among patient demographics, tumor characteristics, treatment regimens, and outcomes, we aimed to identify novel prognostic associations. These findings may support more personalized and effective treatment strategies for breast cancer management, particularly in the context of integrated multimodal therapies.

3. Methods

3.1. Data Collection

This cross-sectional study analyzed clinical data from 216 patients with non-metastatic breast cancer referred to the Radiotherapy Center of Towhid Hospital in Sanandaj, Iran, between 2016 and 2023. Patients with a confirmed diagnosis and complete clinical and treatment records were included. The study was approved by the Ethics Committee of Kurdistan University of Medical Sciences (IR.MUK.REC.1399/064) and conducted in accordance with the Declaration of Helsinki. Data were extracted from medical records using a structured checklist. The variables included the following:

Demographics: Age

2. Tumor characteristics: Primary tumor size, histologic/pathology type, pathology grade, and hormone receptor status (ER, PR, and HER2). Nodal status (N stage) was not available in a consistently extractable format across the medical records and was therefore not included in the analyses.

3. Treatment details: First-line treatment modality (surgery, chemotherapy, or radiotherapy), whether each modality was performed, type of surgery (lumpectomy vs mastectomy), and prescribed radiotherapy dose/fractionation, as recorded in radiotherapy treatment charts. Radiotherapy prescriptions at our center followed institutional practice during the study period. Whole-breast or chest-wall irradiation was most commonly delivered using conventional fractionation of 45 - 50.4 Gy in 1.8 - 2.0 Gy per fraction, with a 10 - 16 Gy tumor-bed boost for high-risk margins when indicated. Hypofractionated schedules, such as 40 - 42.5 Gy in 15 - 16 fractions, were used selectively for patients judged eligible based on clinical risk and physician preference; however, hypofractionation flags were not systematically recorded in older records. Detailed radiotherapy target volumes, such as explicit chest-wall vs regional nodal fields, were not consistently coded in the electronic dataset and therefore were not used as independent variables in the present analysis.

4. Outcome: Survival status, defined as alive or dead at the end of 2023.

3.2. Data Preprocessing

All preprocessing was performed in Python using pandas and standard scientific libraries. The aim was to produce a reliable, consistently encoded transaction-style dataset suitable for ARM while minimizing biased signals introduced by inconsistently recorded variables.

3.2.1. Cleaning, Standardization, and Exclusion of Unreliable Variables

Records with missing essential data, including survival status, treatment indicators, or receptor status, were removed ($n = 18$, 8.3%). Categorical inconsistencies, such as "Alive" vs "alive," were standardized. Variables with substantial missing data due to inconsistent documentation were excluded to avoid spurious associations. These variables included nodal stage (100% missing, free-text only), chemotherapy regimens (89% missing, inconsistent drug/cycle recording), radiotherapy target volumes (100% missing, not coded), fractionation schedules (76% missing, not systematically documented), endocrine therapy (96% missing,

externally managed), and trastuzumab receipt (94% missing, referrals unconfirmed). These variables were excluded rather than imputed because missingness was non-random. This approach prioritized data integrity while acknowledging residual confounding as a limitation.

3.2.2. Variable Transformation and Encoding

To enable rule mining, all continuous variables were converted into clinically meaningful categorical bins, as summarized in Table 1. Prescribed total radiotherapy dose was used as a pragmatic proxy for radiotherapy exposure because fractionation schedules were not systematically recorded.

3.2.3. Transaction Formatting for Association Rule Mining

Each patient record was converted into a transaction consisting of item-like entries, such as "Age > 60," "HER2 Negative," "Received Chemotherapy," and "Alive." This transaction-style format was required to extract meaningful association rules using ARM algorithms. These preprocessing steps ensured that the dataset was clean, consistent, and ready for mining meaningful patterns related to tumor characteristics, treatment decisions, and survival outcomes in patients with breast cancer.

3.3. Association Rule Mining

Association rule mining enables the identification of clinically relevant patterns that may influence survival and treatment outcomes, thereby supporting clinical decision-making. The Apriori algorithm, implemented in Python, was used to derive these associations. The algorithm works in 3 main steps: 1) identifying frequent single items that meet the support threshold, 2) combining them into larger frequent itemsets, and 3) deriving rules that satisfy the confidence and lift criteria. The Apriori algorithm was used to identify frequent combinations of clinical features and treatments associated with survival. Rules are expressed as antecedent $\rightarrow$ consequent, such as {Tumor Size > 5 cm, Received Chemotherapy} $\rightarrow$ {Alive}. Rule quality was assessed using the following standard metrics:

1. Support: Frequency of the itemset in the data.
2. Confidence: Conditional probability of the consequent given the antecedent.
3. Lift: The extent to which the consequent is more likely given the antecedent. A lift > 1 indicates a positive association.

To ensure clinical relevance and statistical robustness, we applied thresholds of support $\geq 5\%$ and

Table 1. Summary of Variable Transformations and Encodings

Variables	Original Type	Transformation
Age (y)	Continuous	Categorized as ≤ 40 , 41 - 50, 51 - 60, and > 60 years
Tumor size (cm)	Continuous	Categorized as 0.5 - 2 (small), 2 - 5 (medium), and > 5 (large)
Prescribed dose (Gy)	Continuous	Coded as no radiotherapy, < 45 (low), 45 - 50.4 (standard), 50.5 - 60 (moderate-high), and > 60 (high)
ER, PR, and HER2	Binary	Standardized as positive or negative
Survival outcome	Binary	Standardized as alive or dead
Surgery, radiotherapy, and chemotherapy	Categorical	Coded as yes or no to indicate whether treatment was received

confidence $\geq 60\%$ and retained only rules with lift > 1 . The top 10 rules by support were selected for interpretation. Figure 1 illustrates the overall analytical workflow. To assess rule robustness given the modest sample size, we performed sensitivity analyses using varying minimum support thresholds (3%, 5%, 7%, and 10%) while maintaining confidence $\geq 60\%$ and lift > 1 . Rules that persisted at higher support thresholds were considered more robust. In addition, for each rule presented in the Results section, we calculated and reported the absolute number of patients supporting the rule, defined as the raw count of transactions containing both the antecedent and the consequent. This transparency allows readers to assess rule stability based on subgroup sizes.

3.4. Statistical Analysis

To complement rule mining, we assessed monotonic relationships between all clinical variables and five-year survival status. Given the ordinal and non-normally distributed nature of the data, Spearman rank correlation was used. A correlation heatmap was generated in Python using seaborn to visualize the strength and direction of these associations, supporting the identification of feature clusters and potential multicollinearity.

4. Results

This study initially analyzed a cohort of 216 patients with non-metastatic breast cancer treated at Towhid Hospital in Sanandaj, Iran, between 2016 and 2023. After preprocessing and excluding records with incomplete or missing essential variables, 198 patients with complete records remained for ARM analysis. This exclusion may introduce selection bias if missingness was non-random. Table 2 summarizes key characteristics of the final cohort, including demographic variables such as age; tumor features such as size, grade, and pathology; treatment modalities such as surgery, chemotherapy, and radiotherapy; receptor status including HER2, ER, and PR; and survival

outcomes. Survival rates across subgroups are also reported to highlight potential outcome disparities and guide further pattern discovery.

4.1. Association Rules Predicting Five-Year Mortality

Using the Apriori algorithm implemented in Python with a minimum support of 5%, minimum confidence of 60%, and lift > 1 , 10,000 rules were generated from the full breast cancer dataset. Of these, 7306 rules met all 3 thresholds: support ≥ 0.05 , confidence ≥ 0.60 , and lift > 1 . The 10 rules with the highest support are presented below and grouped according to their consequents (“Survival = Dead” vs “Survival = Alive”) to highlight the most frequent predictors of five-year mortality and survival, respectively. Before presenting the rules, we emphasize that ARM identifies co-occurrence, not causation. The associations below reflect patterns in the data that may be influenced by unmeasured confounders, selection bias, or local practices. These patterns do not imply causation and should not be used for clinical decision-making without prospective validation.

Tables 3 and 4 present the top 10 rules by support, including the absolute number of patients supporting each rule. Given the modest sample size, we consider rules supported by fewer than 20 patients to require particularly cautious interpretation because they may be unstable or overfitted. Sensitivity analyses across support thresholds (3%, 5%, 7%, and 10%) showed that rules predicting mortality among HER2-positive patients persisted at higher thresholds (7% - 10%), indicating relative robustness. However, rules involving highly specific combinations, particularly the grade 3, HER2-negative, PR-positive survival profile, did not persist above 5% support, confirming their dependence on smaller subgroups.

4.1.1. Rules Predicting Five-Year Mortality (Survival = Dead)

Table 3 shows the top 10 rules associated with a higher likelihood of death within 5 years. In every rule,

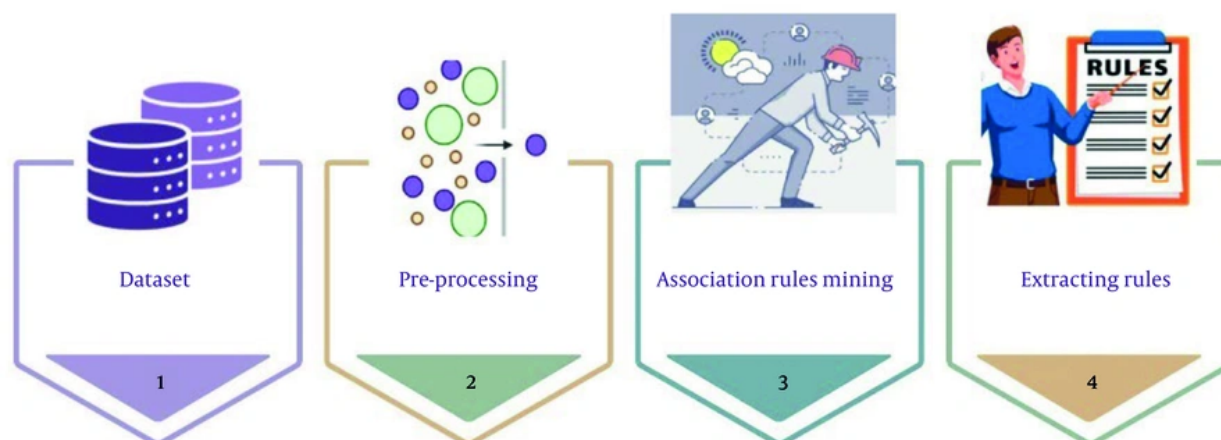


Figure 1. Association rule mining process for exploring meaningful relationships among variables in the dataset

the tumor type was invasive ductal carcinoma, and most rules involved HER2-positive status. For example, rules 1 and 2 (support = 0.061, confidence = 0.706, lift = 2.118) indicate that patients with HER2-positive invasive ductal carcinoma who received standard-dose chemotherapy, regardless of whether they also received radiotherapy, were approximately 2.1 times more likely to die within 5 years than the average patient. Beyond HER2-positive cases, rule 10 (support = 0.051, confidence = 0.625, lift = 1.875) identifies another high-risk group: patients with medium-sized tumors (2 - 5 cm), grade 2 pathology, and surgery as the first treatment, mostly mastectomy, followed by chemotherapy. In other words, even patients with moderately large and moderately aggressive tumors had worse outcomes when treated with surgery plus chemotherapy.

4.1.2. Rules Predicting Five-Year Survival (Survival = Alive)

Table 4 reveals 2 clear low-risk profiles for five-year survival. The first profile included patients aged 50 - 69 years with small (< 2 cm), PR-positive tumors and negative surgical margins. Seven rules, including rules 4 - 9, describe variations of this profile, all with perfect confidence (100% survival) and a lift of 1.5, indicating that these patients were 1.5 times more likely to survive than the average patient. This favorable outcome was independent of whether they received chemotherapy or radiotherapy. The second profile consisted of patients aged 50 - 69 years with grade 3, HER2-negative, PR-positive invasive ductal carcinoma, as shown in rules 1 - 3. Despite high-grade histology, these patients also

achieved 100% five-year survival (lift = 1.5) when treated with mastectomy or standard-dose radiotherapy. This finding suggests that hormone receptor positivity and HER2 negativity can mitigate the prognostic impact of high tumor grade in this age group. Rule 10 (support = 0.071, confidence = 0.947, lift = 1.421) further reinforces the excellent prognosis associated with small, margin-negative tumors in patients aged 50 - 69 years, even without considering receptor status.

4.2. Correlation Analysis

To complement rule mining, we assessed monotonic relationships between clinical variables and five-year survival using Spearman rank correlation (Figure 2). Categorized tumor size showed the strongest negative correlation with survival ($\rho = -0.14$), confirming that larger tumors were associated with poorer outcomes. Receipt of chemotherapy showed a modest positive correlation ($\rho = 0.10$), suggesting a survival benefit. Surgery was negatively correlated with survival ($\rho = -0.15$), likely reflecting that sicker patients underwent more aggressive surgery rather than indicating a detrimental effect of surgery itself. Strong intercorrelations were observed between ER and PR ($\rho = 0.75$), surgery and surgery type ($\rho = -0.57$), and radiotherapy and dose range ($\rho = 0.48$), indicating redundancy among these variables. These findings informed feature selection for ARM, in which we prioritized nonredundant and clinically informative variables, such as tumor size and receptor status, to ensure interpretable rules.

Table 2. Characteristics of the Breast Cancer Dataset (N = 198)^a

Characteristics	Frequency	Percentage	Additional Metrics
Survival status			
Alive	132	66.7	
Dead	66	33.3	
Age range			
< 40	32	16.2	Mean age: 48.5 years (SD: 12.3)
40 - 49	79	39.9	
50 - 69	84	42.4	
≥ 70	4	2.0	
Tumor size (categorized)			
Small (< 2 cm)	53	26.8	Mean tumor size: 3.8 cm (SD: 2.2)
Medium (2 - 5 cm)	112	56.6	
Large (> 5 cm)	33	16.7	
Pathology grade			
1	21	10.6	Survival rate (grade 1): 71.4% (15/21)
2	119	60.1	Survival rate (grade 2): 68.1% (81/119)
3	58	29.3	Survival rate (grade 3): 60.3% (35/58)
Pathology type			
Invasive ductal carcinoma	158	79.8	Survival rate: 65.8% (104/158)
Infiltrating ductal carcinoma	23	11.6	Survival rate: 69.6% (16/23)
In situ and invasive ductal carcinoma	6	3.0	Survival rate: 66.7% (4/6)
Invasive lobular carcinoma	11	5.6	Survival rate: 72.7% (8/11)
Surgery			
Yes	180	90.9	Survival rate (yes): 67.2% (121/180)
No	18	9.1	Survival rate (no): 61.1% (11/18)
Surgery type			
Mastectomy	117	59.1	Survival rate: 65.8% (77/117)
Lumpectomy	63	31.8	Survival rate: 69.8% (44/63)
No surgery	18	31.8	Survival rate: 69.8% (44/63)
Received radiotherapy			
Yes	181	91.4	Survival rate (yes): 66.3% (120/181)
No	17	8.6	Survival rate (no): 70.6% (12/17)
Dose range (Gy)			
Low (< 40)	19	9.6	Survival rate: 63.2% (12/19)
Standard (45 - 50.5)	155	78.3	Survival rate: 67.1% (104/155)
Moderate-high (> 50.4)	6	3.0	Survival rate: 66.7% (4/6)
No radiotherapy	17	8.6	Survival rate: 70.6% (12/17)
Received chemotherapy			
Yes	106	53.5	Survival rate (yes): 64.2% (68/106)
No	92	46.5	Survival rate (no): 69.6% (64/92)
HER2 status			
Positive	67	33.8	Survival rate (positive): 61.2% (41/67)
Negative	131	66.2	Survival rate (negative): 69.5% (91/131)
ER status			
Positive	158	79.8	Survival rate (positive): 67.1% (106/158)
Negative	40	20.2	Survival rate (negative): 65.0% (26/40)
PR status			
Positive	153	77.3	Survival rate (positive): 66.7% (102/153)
Negative	45	22.7	Survival rate (negative): 66.7% (30/45)

^a Abbreviation: SD, Standard deviation.

5. Discussion

In this study, we applied ARM to a cohort of 198 patients with non-metastatic breast cancer treated with multimodal therapy to identify clinically interpretable

Table 3. Top Ten Association Rules (Support $\geq 5\%$, Confidence $\geq 60\%$, Lift > 1) Predicting Five-Year Mortality (Survival = Dead)

Antecedent (X)	Support	Confidence	Lift	N Supporting ^a
Chemotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, Dose Range = Standard, HER2 = Positive	0.061	0.706	2.118	12
Radiotherapy = Yes, Chemotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, Dose Range = Standard, HER2 = Positive	0.061	0.706	2.118	12
Radiotherapy = Yes, Chemotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, HER2 = Positive	0.066	0.684	2.053	13
Chemotherapy = Yes, ER = Positive, Pathology Type = Invasive Ductal Carcinoma, Dose Range = Standard, HER2 = Positive	0.051	0.667	2.000	10
Radiotherapy = Yes, Chemotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, Surgery = Yes, HER2 = Positive	0.051	0.667	2.000	10
Radiotherapy = Yes, Chemotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, HER2 = Positive, PR = Positive	0.051	0.667	2.000	10
Radiotherapy = Yes, Chemotherapy = Yes, ER = Positive, Pathology Type = Invasive Ductal Carcinoma, HER2 = Positive	0.056	0.647	1.941	14
Chemotherapy = Yes, Dose Range = Standard, HER2 = Positive	0.071	0.636	1.909	14
Radiotherapy = Yes, Chemotherapy = Yes, Dose Range = Standard, HER2 = Positive	0.071	0.636	1.909	10
Tumor Size Categorized = Medium, Chemotherapy = Yes, Pathology Grade = 2, First Treatment = Surgery, Surgery Type = Mastectomy	0.051	0.625	1.875	10

^a N Supporting: absolute number of patients out of 198 who exhibited both the antecedent conditions and the consequent (Survival = Dead). Values are rounded to the nearest integer. Rules with N < 20 should be interpreted cautiously because of potential instability.

Table 4. Top Ten Association Rules (Support $\geq 5\%$, Confidence $\geq 60\%$, Lift > 1) Predicting Five-Year Survival (Survival = Alive)

Antecedent (X)	Support	Confidence	Lift	N Supporting
Pathology Type = Invasive Ductal Carcinoma, Pathology Grade = 3, HER2 = Negative, PR = Positive, Surgery Type = Mastectomy	0.066	1.000	1.500	13
Radiotherapy = Yes, Pathology Type = Invasive Ductal Carcinoma, Pathology Grade = 3, Age Range = 50 - 69, HER2 = Negative	0.056	1.000	1.500	11
Pathology Type = Invasive Ductal Carcinoma, Pathology Grade = 3, Age Range = 50 - 69, Dose Range = Standard, HER2 = Negative	0.056	1.000	1.500	11
Tumor Size Categorized = Small, Pathology Type = Invasive Ductal Carcinoma, Surgery Margin = Negative, Age Range = 50 - 69, PR = Positive	0.071	1.000	1.500	14
Tumor Size Categorized = Small, Surgery = Yes, Surgery Margin = Negative, Age Range = 50 - 69, PR = Positive	0.076	1.000	1.500	15
Tumor Size Categorized = Small, Surgery Margin = Negative, Age Range = 50 - 69, First Treatment = Surgery, PR = Positive	0.071	1.000	1.500	14
Tumor Size Categorized = Small, Surgery Margin = Negative, Age Range = 50 - 69, Dose Range = Standard, PR = Positive	0.071	1.000	1.500	14
Tumor Size Categorized = Small, Chemotherapy = Yes, Age Range = 50 - 69, Dose Range = Standard	0.051	1.000	1.500	10
Tumor Size Categorized = Small, Radiotherapy = Yes, Chemotherapy = Yes, Age Range = 50 - 69, Dose Range = Standard	0.051	1.000	1.500	10
Tumor Size Categorized = Small, Surgery = Yes, Surgery Margin = Negative, Age Range = 50 - 69	0.071	0.947	1.421	14

patterns associated with five-year survival. Our analysis revealed distinct prognostic profiles: HER2-positive invasive ductal carcinoma treated with standard-dose radiotherapy and/or chemotherapy emerged as a high-risk pattern, whereas patients aged 50 - 69 years with small, PR-positive, margin-negative tumors, as well as an unexpectedly favorable subset with grade 3, HER2-negative, PR-positive disease, demonstrated excellent outcomes. Although these findings are associative, they generate hypotheses about how tumor biology and patient characteristics interact with treatment modalities to influence survival.

Our findings are broadly consistent with previous literature emphasizing the prognostic importance of tumor burden, receptor status, and multimodal therapy (30). The high-risk profile associated with HER2-positive disease aligns with large-scale analyses showing that

HER2 positivity confers increased recurrence risk (31) and can be associated with resistance to standard chemotherapy regimens in the absence of targeted therapy (12). Our finding that this risk persisted despite receipt of standard radiotherapy and chemotherapy underscores the critical importance of anti-HER2 agents, such as trastuzumab, which were not systematically recorded in our dataset. This omission likely explains the elevated mortality observed in this subgroup and highlights a potential gap in guideline-concordant care that warrants further investigation.

The favorable prognosis observed in patients with small, PR-positive, margin-negative tumors is well supported by the literature. More notably, our identification of a low-risk profile among patients aged 50 - 69 years with grade 3, HER2-negative, PR-positive tumors adds nuance to the understanding of high-grade

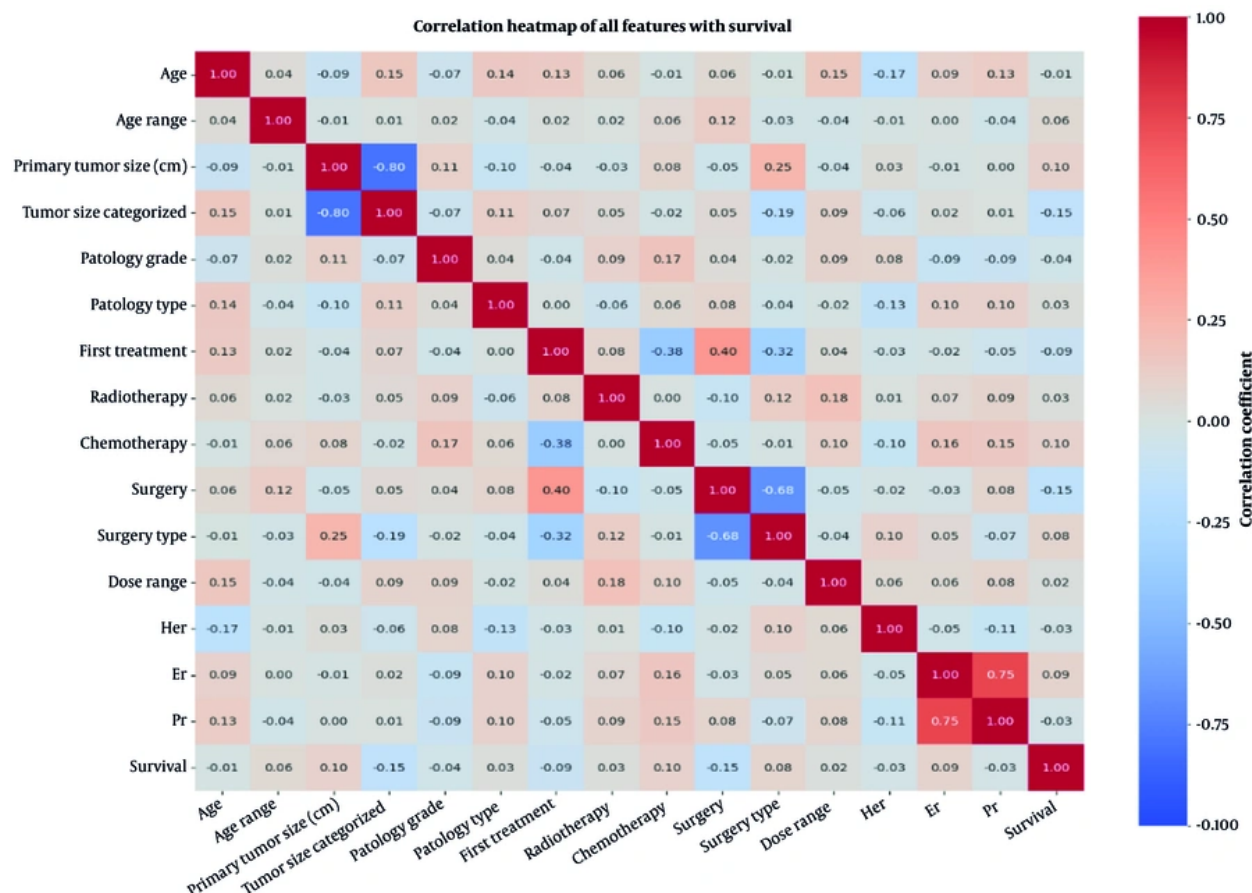


Figure 2. Heatmap of pearson correlation coefficients for clinical features and five-year survival in breast cancer patients

disease. Although grade 3 histology is typically associated with aggressive behavior, our results suggest that, in the context of hormone receptor positivity, HER2 negativity, and appropriate locoregional therapy, outcomes may be excellent. This finding echoes subtype-adjusted analyses showing that hormone receptor-positive, HER2-negative cancers have favorable prognoses even when other high-risk features are present (31). The correlation analyses reinforced the centrality of tumor size as a prognostic factor and revealed expected collinearities, such as ER/PR and surgery/surgery type, which informed our feature selection for ARM and ensured rule interpretability.

Figure 3 illustrates the hypothesis-generating risk patterns identified by ARM. HER2-positive patients had higher mortality despite standard treatment, underscoring the need for early targeted therapies such

as trastuzumab. Patients with medium-sized, grade 2 tumors treated with mastectomy and chemotherapy remained at risk; closer follow-up or genomic profiling might be helpful in this subgroup. Two low-risk groups performed exceptionally well: women aged 50 - 69 years with small, PR-positive, margin-negative tumors and those with grade 3, HER2-negative, PR-positive disease. For the first group, shorter radiation or endocrine therapy alone could reduce toxicity. For the second group, multiagent chemotherapy might be unnecessary. High-risk patients may require more intensive monitoring, whereas low-risk patients could shift to less frequent follow-up focused on quality of life. These hypotheses require prospective validation before clinical implementation.

5.1. Strengths and Novel Contributions

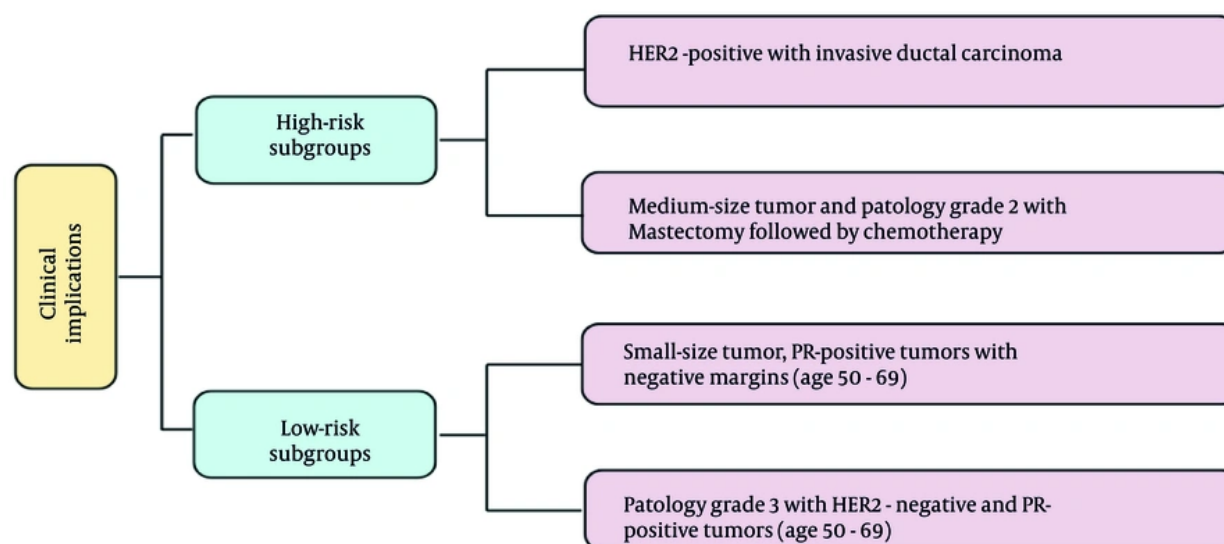


Figure 3. Clinical implications of risk-stratified subgroups in early-stage breast cancer

This study offers several strengths and novel contributions to breast cancer prognosis. Unlike many ARM studies that focus only on demographics or baseline tumor features (29), this work integrates treatment variables with tumor biology within an ARM framework and generates directly interpretable rules that complement traditional survival analyses. This hypothesis-generating approach can reveal patterns that prompt further investigation, such as the unexpected finding that a subset of patients with grade 3 disease had excellent outcomes. Furthermore, by focusing on a uniformly treated real-world cohort from a single center, we minimized practice variation that can confound larger registry-based analyses.

5.2. Limitations

Despite these strengths, several important limitations should be considered. First and most importantly, ARM identifies co-occurrence rather than causation. Our rules represent statistical associations that may be driven by unmeasured confounders, such as nodal stage, comorbidity, performance status, and socioeconomic factors, as well as selection biases or local treatment practices. All findings are therefore hypothesis-generating, and clinical implications, such as treatment de-escalation or intensified surveillance, are presented only as hypotheses to be tested in future

interventional studies, not as recommendations. The single-center retrospective design and exclusion of 18 patients with incomplete records may introduce selection bias, thereby limiting external validity. Multicenter validation is essential before any clinical translation.

We lacked consistently extractable data on nodal staging, detailed chemotherapy regimens, systematic fractionation or target-volume coding, and reliable records of endocrine or HER2-targeted therapy receipt; these omissions limited our ability to account for key prognostic and treatment modifiers. Total prescribed dose categories served as a pragmatic proxy for radiotherapy exposure, although different fractionation schedules with equivalent total doses have different biological effects. The dataset exhibited moderate class imbalance, with 66.7% alive vs 33.3% dead, which may have biased rule discovery toward the majority class, as ARM inherently favors frequent itemsets. We did not perform balanced sampling or correction techniques because our primary goal was hypothesis generation using the full real-world cohort, and any correction method would introduce its own biases given our modest sample size. Therefore, minority-class patterns, particularly the grade 3, HER2-negative, PR-positive survival profile, should be interpreted with particular caution. Finally, some subgroup rules are based on small numbers and therefore carry wide uncertainty despite

high apparent confidence; readers should avoid overinterpretation.

5.3. Future Directions

Several steps are needed to increase robustness and generalizability: 1) validating ARM-derived rules in larger external cohorts with complete nodal staging and treatment data; 2) applying survival-specific methods, such as survival trees and time-dependent ARM, that can accommodate censored data; 3) addressing class imbalance through balanced sampling or synthetic minority oversampling in larger datasets; 4) integrating genomic signatures, such as Oncotype DX, to enhance biological plausibility; and 5) prospectively evaluating whether rule-based risk stratification improves clinical decision-making. We are actively collecting a multicenter dataset to address these goals. If validated, ARM rules could be incorporated into decision-support prototypes that present interpretable risk patterns alongside traditional survival models.

5.4. Conclusions

In summary, applying ARM to routine clinical data revealed clear multivariable patterns that are consistent with established prognostic factors, such as tumor size, receptor status, and treatment approach, and that point to possible high- and low-risk patient groups. Some of these patterns confirmed well-known relationships, whereas others, such as excellent short-term survival in certain older patients with grade 3, HER2-negative, PR-positive tumors, were less expected and merit further study. These findings are hypothesis-generating and should be validated in larger, prospective, multicenter studies. With the addition of more detailed clinical, genomic, and radiomic data, ARM-derived rules could be integrated into decision-support tools to help guide personalized treatment strategies and potentially improve patient outcomes.

Footnotes

AI Use Disclosure: The authors declare that no generative AI tools were used in the creation of this article.

Authors' Contribution: Study concept/design: H. G. and F. A.; Data acquisition: F. A.; Data analysis/interpretation: H. G., J. A., and H. P.; Manuscript drafting: H. G. and J. A.; Critical revision for important intellectual content: All authors; Statistical analysis: H. P.; Administrative/technical/material support: All authors; Study supervision: H. G.

Conflict of Interests Statement: The authors do not declare any conflicts of interests for this study.

Data Availability: The dataset presented in the study is available on request from the corresponding author during submission or after publication.

Ethical Approval: This study is a retrospective analysis, and all data were obtained from patient records at The Tohid Hospital of Kurdistan University of Medical Sciences. The research protocol was reviewed and approved by the Ethics Committee of Kurdistan University of Medical Sciences under the approval code IR.MUK.REC.1399/064. All methods were carried out in accordance with relevant guidelines and regulations. This study adheres to the ethical principles outlined in the Declaration of Helsinki and its subsequent amendments, ensuring the protection of patient privacy and data confidentiality.

Funding/Support: No funding was received for this study.

References

- Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229-63. [PubMed ID: 38572751]. <https://doi.org/10.3322/caac.21834>.
- Zhu Z, Jiang L, Ding X. Advancing breast cancer heterogeneity analysis: Insights from genomics, transcriptomics and proteomics at bulk and single-cell levels. *Cancers.* 2023;15(16):4164. [PubMed ID: 37627192]. [PubMed Central ID: PMC10452610]. <https://doi.org/10.3390/cancers15164164>.
- Parikh D, Shah M. A comprehensive study on epigenetic signatures to monitor disease progression and the response to therapy in breast cancer. *Biomedical Analysis.* 2024;1(3):205-17. <https://doi.org/10.1016/j.bioana.2024.06.004>.
- MacDonald WJ, Purcell C, Pinho-Schwermann M, Stubbs NM, Srinivasan PR, El-Deiry WS. Heterogeneity in Cancer. *Cancers.* 2025;17(3):441. [PubMed ID: 39941808]. [PubMed Central ID: PMC11816170]. <https://doi.org/10.3390/cancers17030441>.
- Najafi S, Sadeghi M, Abasvandi F, Shajari MR, Mohebi K, Ghandchi H. Prognostic factors influencing prognosis in early breast cancer patients. *Menopause Review/Przegląd Menopauzalny.* 2019;18(2):82-8. [PubMed ID: 31485204]. [PubMed Central ID: PMC6719640]. <https://doi.org/10.5114/pm.2019.86833>.
- Łukasiewicz S, Czeczulewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. *Cancers.* 2021;13(17):4287. [PubMed ID: 34503097]. [PubMed Central ID: PMC8428369]. <https://doi.org/10.3390/cancers13174287>.
- Zheng F, Du F, Yang Z, Wang X, Yue J, Ling Y, et al. Potential prognostic value of HER2/CEP17 FISH ratio in HER2-positive non-metastatic breast cancer: A real-world study. *Discover Oncology.* 2025;16(1): 800. [PubMed ID: 40382749]. [PubMed Central ID: PMC12086127]. <https://doi.org/10.1007/s12672-025-02495-2>.
- Zumsteg ZS, Morrow M, Arnold B, Zheng J, Zhang Z, Robson M, et al. Breast-conserving therapy achieves locoregional outcomes

- comparable to mastectomy in women with T1 - 2N0 triple-negative breast cancer. *Ann Surg Oncol*. 2013;**20**(11):3469-76. [PubMed ID: 23686101]. [PubMed Central ID: PMC5730455]. <https://doi.org/10.1245/s10434-013-3011-9>.
9. van Maaren MC, de Munck L, Strobbe LJA, Sonke GS, Westenend PJ, Smidt ML, et al. Ten-year recurrence rates for breast cancer subtypes in the Netherlands: A large population-based study. *Int J Cancer*. 2019;**144**(2):263-72. [PubMed ID: 30368776]. <https://doi.org/10.1002/ijc.31914>.
 10. Mandapati A, Lukong KE. Triple negative breast cancer: Approved treatment options and their mechanisms of action. *J Cancer Res Clin Oncol*. 2023;**149**(7):3701-19. [PubMed ID: 35976445]. [PubMed Central ID: PMC10314854]. <https://doi.org/10.1007/s00432-022-04189-6>.
 11. Carvalho FM. Targeting low-risk triple-negative breast cancer: A review on de-escalation strategies for a new era. *Transl Breast Cancer Res*. 2024;**6**:4-4. [PubMed ID: 39980807]. [PubMed Central ID: PMC11836748]. <https://doi.org/10.21037/tbcr-24-28>.
 12. Blum JL, Flynn PJ, Yothers G, Asmar L, Geyer CE, Jacobs SA, et al. Anthracyclines in early breast cancer: The ABC trials-USOR 06 - 090, NSABP B-46-I/USOR 07132, and NSABP B-49 (NRG Oncology). *J Clin Oncol*. 2017;**35**(23):2647-55. [PubMed ID: 28398846]. [PubMed Central ID: PMC5549453]. <https://doi.org/10.1200/JCO.2016.71.4147>.
 13. Nitz U, Gluz O, Clemens M, Malter W, Reimer T, Nuding B, et al. West German Study PlanB trial: Adjuvant 4 cycles of epirubicin and cyclophosphamide plus docetaxel versus 6 cycles of docetaxel and cyclophosphamide in HER2-negative early breast cancer. *J Clin Oncol*. 2019;**37**(10):799-808. [PubMed ID: 30785826]. <https://doi.org/10.1200/JCO.18.00028>.
 14. Poggio F, Bruzzone M, Ceppi M, Pondé NF, La Valle G, Del Mastro L, et al. Platinum-based neoadjuvant chemotherapy in triple-negative breast cancer: A systematic review and meta-analysis. *Ann Oncol*. 2018;**29**(7):1497-508. [PubMed ID: 29873695]. <https://doi.org/10.1093/annonc/mdy127>.
 15. von Minckwitz G, Schneeweiss A, Loibl S, Salat C, Denkert C, Rezai M, et al. Neoadjuvant carboplatin in patients with triple-negative and HER2-positive early breast cancer (GeparSixto; GBG 66): A randomised phase 2 trial. *Lancet Oncol*. 2014;**15**(7):747-56. [PubMed ID: 24794243]. [PubMed Central ID: PMC1326012]. [https://doi.org/10.1016/S1470-2045\(14\)70160-3](https://doi.org/10.1016/S1470-2045(14)70160-3).
 16. Coles CE, Griffin CL, Kirby AM, Titley J, Agrawal RK, Alhasso A, et al. Partial-breast radiotherapy after breast conservation surgery for patients with early breast cancer (UK IMPORT LOW trial): 5-year results from a multicentre, randomised, controlled, phase 3, non-inferiority trial. *Lancet*. 2017;**390**(10099):1048-60. [PubMed ID: 28779963]. [PubMed Central ID: PMC5594247]. [https://doi.org/10.1016/S0140-6736\(17\)31145-5](https://doi.org/10.1016/S0140-6736(17)31145-5).
 17. Marta GN, Macedo CR, Carvalho HDA, Hanna SA, da Silva JLF, Riera R. Accelerated partial irradiation for breast cancer: Systematic review and meta-analysis of 8653 women in eight randomized trials. *Radiother Oncol*. 2015;**114**(1):42-9. [PubMed ID: 25480094]. <https://doi.org/10.1016/j.radonc.2014.11.014>.
 18. Garg P, Goyal N, Priya S, Grover R, Singh S. Hypofractionated radiotherapy in carcinoma breast-A 5-year single institution experience. *J Cancer Res Ther*. 2025;**21**(5):995-9. [PubMed ID: 41004287]. https://doi.org/10.4103/jcrt.jcrt_282_24.
 19. McGale P, Correa C, Cutter D, Duane F, Ewertz M, Gray R, et al. Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: Meta-analysis of individual patient data for 8135 women in 22 randomised trials. *Lancet*. 2014;**383**(9935):2127-35. [PubMed ID: 24656685]. [PubMed Central ID: PMC5015598]. [https://doi.org/10.1016/S0140-6736\(14\)60488-8](https://doi.org/10.1016/S0140-6736(14)60488-8).
 20. Whelan TJ, Olivetto IA, Parulekar WR, Ackerman I, Chua BH, Nabid A, et al. Regional nodal irradiation in early-stage breast cancer. *N Engl J Med*. 2015;**373**(4):307-16. [PubMed ID: 26200977]. [PubMed Central ID: PMC4556358]. <https://doi.org/10.1056/NEJMoa1415340>.
 21. Hennequin C, Bossard N, Servagi-Vernat S, Maingon P, Dubois JB, Datchary J, et al. Ten-year survival results of a randomized trial of irradiation of internal mammary nodes after mastectomy. *Int J Radiat Oncol Biol Phys*. 2013;**86**(5):860-6. [PubMed ID: 23664327]. <https://doi.org/10.1016/j.ijrobp.2013.03.021>.
 22. Obeagu EI, Obeagu GU. Predictive models and biomarkers for survival in stage III breast cancer: A review of clinical applications and future directions. *Ann Med Surg*. 2024;**86**(10):5980-7. [PubMed ID: 39359789]. [PubMed Central ID: PMC11444610]. <https://doi.org/10.1097/MS9.0000000000002517>.
 23. Sammut SJ, Crispin-Ortuzar M, Chin SF, Provenzano E, Bardwell HA, Ma W, et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*. 2022;**601**(7894):623-9. [PubMed ID: 34875674]. [PubMed Central ID: PMC8791834]. <https://doi.org/10.1038/s41586-021-04278-5>.
 24. Price JM, Prabhakaran A, West CML. Predicting tumour radiosensitivity to deliver precision radiotherapy. *Nat Rev Clin Oncol*. 2023;**20**(2):83-98. [PubMed ID: 36477705]. <https://doi.org/10.1038/s41571-022-00709-y>.
 25. Bleaney CW, Abdelal H, Reardon M, Anandadas C, Hoskin P, Choudhury A, et al. Clinical biomarkers of tumour radiosensitivity and predicting benefit from radiotherapy: A systematic review. *Cancers*. 2024;**16**(10):1942. [PubMed ID: 38792019]. [PubMed Central ID: PMC1119069]. <https://doi.org/10.3390/cancers16101942>.
 26. Shu X, Ye Y. Knowledge discovery: Methods from data mining and machine learning. *Soc Sci Res*. 2023;**110**: 102817. [PubMed ID: 36796993]. <https://doi.org/10.1016/j.ssres.2022.102817>.
 27. Verma S, Vishal, Mittal K, Syed M, Gupta P, Majumdar G, et al. *Association Rule Mining in Reproductive Health Data*. Data-Driven Reproductive Health: Springer; 2024. p. 85-111. https://doi.org/10.1007/978-981-97-7451-7_5.
 28. Hormaty S, Seiwon AN, Rasheed BH, Parvaz H, Gharahzadeh A, Ghaznavi H. A review on biomarker-enhanced machine learning for early diagnosis and outcome prediction in ovarian cancer management. *Cancer Med*. 2025;**14**(17): e71224. [PubMed ID: 40927964]. [PubMed Central ID: PMC12421415]. <https://doi.org/10.1002/cam4.71224>.
 29. Adam N, Wieder R. Temporal association rule mining: Race-based patterns of treatment-adverse events in breast cancer patients using SEER-Medicare dataset. *Biomedicine*. 2024;**12**(6):1213. [PubMed ID: 38927419]. [PubMed Central ID: PMC1200891]. <https://doi.org/10.3390/biomedicine12061213>.
 30. Griem KL, Henderson IC, Gelman R, Ascoli D, Silver B, Recht A, et al. The 5-year results of a randomized trial of adjuvant radiation therapy after chemotherapy in breast cancer patients treated with mastectomy. *J Clin Oncol*. 1987;**5**(10):1546-55. [PubMed ID: 3309196]. <https://doi.org/10.1200/JCO.1987.5.10.1546>.
 31. Noman SM, Fadel YM, Henedak MT, Attia NA, Essam M, Elmaasarawii S, et al. Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis. *Sci Rep*. 2025;**15**(1): 3728. [PubMed ID: 39880868]. [PubMed Central ID: PMC11779859]. <https://doi.org/10.1038/s41598-025-87622-3>.