**Zero : Jurnal Sains, Matematika, dan Terapan**
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 9, Number 3, 2025
DOI: 10.30829/zero.v9i3.26030
Page: 766-780

# Implications of Age-Based Clustering for Survival and Relapse-Free Analysis in METABRIC Breast Cancer

[1] Alif Azhari
Departement of Mathematics, Hasanuddin University, Makassar, 90245, Indonesia

[2] Mauliddin
Departement of Mathematics, Hasanuddin University, Makassar, 90245, Indonesia

Article Info

*Article history:*

Accepted, 26 December 2025

*Keywords:*

Breast Cancer;
Machine Learning;
Relapse-free;
Survival Analysis;
Survival Time.

## ABSTRACT

Cox proportional hazards models are widely used for breast cancer survival analysis, but their validity is often limited by violations of the proportional hazard assumption. Machine learning techniques offer potential ways to improve model robustness, yet their combined use remains underexplored. This study aims to compare the proportional hazard assumptions fulfilment and the discriminatory ability of the models before and after age-based clustering. K-medoids was selected for its robustness to outliers. The results demonstrate that clustering significantly improved adherence to the proportional hazards assumption and increased the concordance index, indicating better predictive performance. Number of variables satisfying the assumption increased from 3 in the global model to 5–6 across clusters. Tumor size and positive lymph nodes consistently had a significant effect on all clusters for both survival time and relapse-free time. These findings suggest that age-based clustering can enhance the robustness and predictive performance of Cox models.

*Corresponding Author:*

Alif Azhari,
Department of Mathematics,
Hasanuddin University, Makassar, Indonesia
Email: alifazharir@gmail.com

## 1. INTRODUCTION

Cancer is a deadly non-communicable disease that can increases mortality risk. According to a report by the International Agency for Research (IARC), breast cancer is the second most common cancer worldwide after lung cancer, with a total incidence of 2,296,840 cases in 2022. When looking only at female gender, breast cancer is the most common cancer affecting women, accounting for 23.8% of all cancer cases worldwide. In 2022, breast cancer accounted for 666,103 deaths, making it the fourth leading cause of cancer mortality globally [1]. Given its high incidence and mortality, identifying prognostic factors for breast cancer survival remains an important research focus. This study aims to analyze significant risk factors for breast cancer patients based on differences in patient age by carefully paying attention to the fulfilment of survival analysis.

Health datasets especially time-to-event data are often difficult to acquire and frequently suffer from high number of missing values. Rather than applying listwise deletion (removing incomplete observations), missing data imputation is a preferable approach as it preserves information from another variable at same sample hence could reduce the potential for biased results. But simple imputation methods such as mean or median imputation can introduce bias as they only replace all missing data points with a single summary statistic, thereby fail to account for inter-variable relationships. In contrast, machine learning method predicts each missing value by leveraging other variables as predictors.

The superior performance of the Random Forest (RF) machine learning model over other classifiers has been documented in multiple health studies. A study by Adiga U et al. for example, employed various classification models on a breast cancer dataset. Their findings identified random forest as the optimal model, yielding the highest accuracy and Area Under Curve (AUC) [2]. Yongxin Li analyzed the prognosis of young breast cancer patients by evaluating samples with an age at diagnosis of 40 years or younger [3]. But, this 40-year threshold for defining young patients was established arbitrarily [3], therefore, this study will utilize cluster analysis for age grouping and adopt the maximum value from the lowest age cluster as the definitive threshold.

In medical research, k-medoids is considered more representative because it selects actual patient data points as cluster representatives (medoids), unlike k-means which uses intra-cluster means that creating artificial cluster centers. A comparative study on clustering tuberculosis indicator data found that k-medoids achieved — on both training and testing data— higher accuracy than k-means clustering and hierarchical clustering methods [4]. Specifically for breast cancer data, a previous study found that k-medoids clustering had a relatively smaller Davies-Bouldin Index (DB-Index) and a relatively larger average silhouette score than k-means clustering for the same number of clusters [5].

Survival analyses, such as the Kaplan-Meier curve and Cox proportional hazards model, are widely used to show the influence of an observed variable on survival time. They are often used because the target variable does not need to meet the distributional assumptions required by methods like the Accelerated Failure Time (AFT) model. Kaplan-Meier is essentially a non-parametric and more flexible method, but the cox proportional hazard model still has the Proportional Hazards (PH) assumption that must be met before interpreting the resulting Hazard Ratio (HR). In a previous study on the risk profiles of breast cancer patients [6], Yuan Gu focused only on observing the survival analysis result of each resulting cluster, but in the current study also clustered the data to constructing a model that satisfied the proportional hazards assumption. Their study [6], however, treated relapse-free time as a predictor variable of survival time, but in this paper relapse-free time expected to be dependent variable like survival time to also consider the risk of patient to be hospitalized or receive treatment. This study only used patient data that could be known from the initial diagnosis, so that the results obtained could be used as a prognosis at the patient's first diagnosis without having to wait for a decision about what therapy the patient would undergo.

Based on the review of previous research, the researchers propose this study, titled " Implications of Age-Based Clustering for Survival and Relapse-Free Analysis in METABRIC Breast Cancer," to complement the statistical findings related to breast cancer in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) dataset. This paper contributes by integrating age-based K-medoids clustering with Random Forest imputation to improve the robustness and discriminatory performance of Cox proportional hazards models for both overall survival and relapse-free survival. This study is expected to enrich the information regarding medical risk factors for breast cancer patients. In addition to medical information, patient risk factors for survival time and relapse-free time can also be used by health insurance underwriters to determine the probability of claims occurrence and for calculating claim reserves[7], [8], [9]. However, this study is limited to the interpretation of statistical results and is separate from clinical analysis by oncology experts. In the future, it is hoped that these statistical results can be further explored clinically by researchers in the medical field, especially in cancer treatment studies.

## 2. RESEARCH METHOD

### 2.1 Random Forest

Leo Breiman introduced the modern Random Forest in 2001 [10] and defined it as a combination of classifier trees $\{h(x,k), k = 1,\dots\}$ where $k$ is a vector of independent and identically distributed random vectors, and each tree casts a single vote for the most popular class for a given input $\boldsymbol{x}$.

### 2.1.1 Margin Function

For an ensemble of classifiers $h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \dots, h_K(\boldsymbol{x})$ where each classifier predicts a class for the same input \$x\$, and the training set is drawn from the random vector distribution (X,Y), the margin function is defined as:

$$mg(\boldsymbol{X},\boldsymbol{Y}) = av_k\, I(h_k(\boldsymbol{X}) = Y) - \max_{j\neq Y} av_k I(h_k(\boldsymbol{X}) = j) \quad (1)$$

where $I(.)$ is the indicator function that has a value of 1 if true but 0 if false, and $av_k$ is the average over all trees $k$. The margin function calculates the difference between the tree that correctly predicts **Y** with tree that incorrectly predicts the **Y** class. The larger the margin, the more confident the model's collective decision that $\boldsymbol{Y}$ is correct. If the margin is negative or zero, it means another class received an equal or greater number of votes than the correct class, indicating a potential misclassification. Mathematically, the margin formula can be detailed as follows:

$$mg(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{K}\sum_{k=1}^{K} I(h_k(\boldsymbol{X}) = Y) - \max_{j\neq Y}\frac{1}{K}\sum_{k=1}^{K} I(h_k(\boldsymbol{X}) = j) \tag{2}$$

The Random Forest algorithm seeks to minimize the generalization error, defined as:

$$PE^{*} = P_{X,Y}(mg(\boldsymbol{X},\boldsymbol{Y}) < 0) \tag{3}$$

The theorem presented by Breiman [10] states that as the number of trees increases, it is almost certain that all sequences $\boldsymbol{\Theta_1}, \dots, \boldsymbol{PE}^{*}$ will converge to:

$$P_{X,Y}(P_{\boldsymbol{\Theta}}(h(\boldsymbol{X},\Theta) = Y) - \max_{j\neq Y} P_{\Theta}(h(\boldsymbol{X},\Theta) = j) < 0) \tag{4}$$

### 2.1.2 Strength dan Correlation

Generalization error has an upper bound calculated through strength and correlation. Strength is described as the average power of each tree in estimating the target variable, formulated as:

$$s = E_{X,Y}[mg(\boldsymbol{X},\boldsymbol{Y})] \tag{5}$$

Meanwhile, the correlation between trees must be minimized so that the ensemble learning function is robust and not just focused on the error of a single or very similar trees. Thus, the average correlation between trees is defined as:

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}[\rho(\Theta,\Theta')\sigma_{\Theta}\sigma_{\Theta'}]}{E_{\Theta,\Theta'}[\sigma_{\Theta}\sigma_{\Theta'}]} \tag{6}$$

Therefore, the upper bound for the generalization error is expressed by the inequality:

$$PE^{*} \leq \frac{\bar{\rho}(1 - s^{2})}{s^{2}} \tag{7}$$

### 2.1.3 Missing Value Imputation

Random Forest as an imputation method is performed by training a model on the observed (complete) data and then applying the model to observations with missing values. For a variable $\boldsymbol{X_s}$ with missing values at entries $\boldsymbol{i}_{mis}^{(s)} \subseteq \{1,2,3, \dots n\}$, the dataset $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \dots, \boldsymbol{X_p})$ is first divided into four parts [11]:

a. $y_{obs}^{(s)}$ : Observed (non-missing) values in variable $\boldsymbol{X_s}$

b. $y_{mis}^{(s)}$ : Missing values in variable $\boldsymbol{X_s}$

c. $x_{obs}^{(s)}$ : Values of variables other than $\boldsymbol{X_s}$ at the same entries as $y_{obs}^{(s)}$

d. $x_{mis}^{(s)}$ : Values of variables other than $\boldsymbol{X_s}$ at the same entries as $y_{mis}^{(s)}$

It should be noted that $x_{mis}^{(s)}$ does not contain missing values; rather, it represents the values of other variables that will be used to predict $y_{mis}^{(s)}$ . Before applying the model to $y_{mis}^{(s)}$, a Random Forest model is first trained using $y_{obs}^{(s)}$ as the response variable, and $x_{obs}^{(s)}$ as the predictor. Imputation is performed iteratively, with the initial matrix denoted as $\boldsymbol{X}_{old}^{imp}$ and the imputation result saved as the matrix $\boldsymbol{X}_{new}^{imp}$. The iteration stops when a stopping criterion $\gamma$ is met, specifically when the difference between the new and previous matrices increases for the first time. The difference for numerical data is defined as [11]:

$$\Delta N = \frac{\sum_{j\in N}\left(\boldsymbol{X}_{new}^{imp} - \boldsymbol{X}_{old}^{imp}\right)^{2}}{\sum_{j\in N}\left(\boldsymbol{X}_{new}^{imp}\right)^{2}} \tag{8}$$

whereas for categorical data, it is defined as:

$$\Delta F = \frac{\sum_{j\in F}\sum_{i=1}^{n} \boldsymbol{I}_{X_{new}^{imp}\neq X_{old}^{imp}}}{\#Missing\ Value} \tag{9}$$

## 2.2 Partitioning Around Medoids

Partitioning Around Medoids (PAM) clustering or also known as k-medoids clustering is a data grouping method based on dissimilarity with all other objects in the cluster that considers the original data as a representation of the cluster instead of using the intra-cluster average. [12], [13], [14]. The k-medoids method selects $k$ initial medoids randomly and then iteratively updates the clusters by assigning the cluster member with the smallest total dissimilarity as the new medoids. The algorithm iterates until the minimum total deviation is achieved, defined as the sum of dissimilarities between each observation $x_c$ and the medoid $m_i$ of its assigned cluster $C_i$[12]. In this study, clustering is performed on a one-dimensional variable (age), and the dissimilarity measure is defined using the one-dimensional Euclidean distance:

$$Total\ Deviation = \sum_{i=1}^{k} \sum_{x_c = C_i} |x_c - m_i| \tag{10}$$

where $x_c$ denotes the age of the $c$-th patient and $m_i$ represent the medoid age of cluster $C_i$.

## 2.3 Elbow Method

The elbow method plot is often used to evaluate and select the optimal number of clusters for k-medoids clustering, as seen in several recent studies [15], [16], [17], [18]. The optimal $k$-cluster is chosen at the point where the Within Sum of Square (WSS) value shows a significant decrease in the plot. The WSS value for each $k$ can be calculated with the formula [17]:

$$\mathrm{WSS} = \sum_{\mathrm{k}=1}^{\mathrm{k}} \sum_{\forall \mathrm{x_i}} \|x_i - C_k\|^2 \tag{11}$$

## 2.4 Kaplan-Meier Estimator

The Kaplan-Meier Curve can display the differences in patient death probabilities between two or more groups over time. According to Elisa & John [19], non-parametric analysis is best performed first, before building a parametric survival model, to gain an overview of the observed survival time data. This is particularly relevant as this study performs clustering, making it important to compare probability estimates between clusters non-parametrically. The Kaplan-Meier estimate is calculated using [20]:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{Y_i}\right) \tag{12}$$

## 2.5 Log-rank Test

To ensure that the Kaplan-Meier curves are genuinely different between groups, the log-rank test is conducted to support the evidence that clustering is significant in the data. The log-rank test compares the observed and expected number of events at each failure time using equal weights [21]. The log-rank test tests $H_0: S_1(t) = S_2(t) = S_3(t)$ against $\mathrm{H_1}$: there is at least one difference, that if $\mathrm{H_1}$ is accepted, it means the clusters explain different survival probabilities over time. It uses the following test statistic:

$$\chi^2 = \sum\nolimits_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i} \tag{13}$$

$$O_i - E_i = \sum\nolimits_{j=1}^{n} (m_{ij} - e_{ij}) \tag{14}$$

$$e_{ij} = \frac{n_{ij}}{\sum_{k=1}^{G} n_{kj}} \left(\sum\nolimits_{k=1}^{G} m_{ij}\right) \tag{15}$$

where $G$ is the number of groups, $O_i$ is the observed value for group $i$, $E_i$ is the expected value for group $i$, $m_{ij}$ is the number of patients who experienced an event in group $i$ at time $t_{(j)}$, $n_{ij}$ is number at risk of $i$ group at time $t_{(j)}$, and $e_{ij}$ is expected value of $i$-group at time $t_{(j)}$.

## 2.6 Cox Proportional Hazard

The Cox Proportional Hazards model can measure the fold-risk (Hazard Ratio) of an individual's survival time shortening for each one-unit increase in a predictor. This method seeks to predict the value $e^{\beta_k}$ (which is hazard ratio) by the following formula:

$$h(t|\boldsymbol{x}) = h_0(t)e^{(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x})} \tag{16}$$

where $h(t)$ is the hazard function at time $t$, $h_0(t)$ is the baseline hazard (the hazard when all covariates are zero). However, Cox proportional hazard has important assumption that is proportional hazard assumption, which assumes that the hazard value of target variable and predictor needs to be constant over time, which means there is no significant change over time in the hazard of all the objects being compared. The survival time of each individual is also assumed to be independent (event that happen in one individual does not affect the other individual's event).

### 2.7 Concordance Index

The Concordance Index (C-Index) is used to evaluate the features used in predicting survival outcomes from a model. The C-Index measures the ability of a model to correctly rank survival times based on predictor values [22], [23]. For example, in the case of predictors and response variables that are directly proportional, then after the response variables are sorted from smallest to largest, the match of the predictor order from smallest to largest is calculated as the C-Index. The C-Index in survival analysis calculated as [24]:

$$\hat{C} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I\left(T_i^{obs} < T_j^{obs}\right) I(M_i > M_j)}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs})} \tag{17}$$

where $T_i^{obs}$ is the time to event or censoring, and $\Delta_i$ is the event indicator (1 if event, 0 if censored). If $\Delta_i = 1$ then $T_i^{obs} = T_i$, in opposite, if $\Delta_i = 0$ then $T_i > T_i^{obs}$ because the event must have occurred after the recording time in the dataset whose exact value is unknown. Moreover $M_i$ is the predicted risk score from the model for $i$ subject. Likewise in random forest formula, $I(.)$ is the indicator function. A pair of values is said to be discordant if $M_i > M_j$ and $T_i^{obs} < T_j^{obs}$ or vice versa. The more discordant pairs, the closer the C-Index is to 0.5 (random chance), while 1.0 is perfect order prediction.

### 2.8 Likelihood Ratio Test

The likelihood ratio (LRT) tests whether all predictors simultaneously have a significant effect on the model (i.e., compares a full model to a reduced/null model) [25]. The LRT testing $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$ over $H_1\colon at\ least\ one\ \beta_J \neq 0$ , using the following statistic [20]:

$$LR = -2\, ln\, L_R - (-2\, ln\, L_F) \sim \chi^2_{\alpha,p} \tag{18}$$

where $p$ is the number of independent variables, $L_R$ is the likelihood for the reduced model (no variables), and $L_F$ is the likelihood for the full model. The LRT can be used as an alternative strategy to evaluate the overall significance of the model [26]. By adding likelihood ratio test can improve the interpretability of the analysis result and increase clinician's confidence in the diagnosis [27].

### 2.9 Wald Test

The Wald test assesses whether individual predictors are significant in the model. It tests $H_0\colon \boldsymbol{R}\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$ against $H_1\colon \boldsymbol{R}\widehat{\boldsymbol{\beta}} \neq \boldsymbol{0}$, using the following statistic [28]:

$$W = \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{0}\right)\left[\boldsymbol{R}\,\widehat{\mathrm{var}(\widehat{\boldsymbol{\beta}})}\,\boldsymbol{R}\right]^{-1}(\boldsymbol{R}\widehat{\boldsymbol{\beta}} - \boldsymbol{0}) \tag{19}$$

where $\widehat{\boldsymbol{\beta}}$ is $p \times 1$ estimation parameter vector of cox regression model, $\boldsymbol{R}$ is hypothesis matrix $p \times q$ size that state the linear combination from parameter tested.

### 2.10 Data and Method

This study uses a quantitative method with time-to-event secondary data obtained from the cBioPortal repository page [29] which combines research published by Curtis C. et al. in Nature 2012 [30] and Pereira et al. in Nature Communications 2016 [31]. The METABRIC dataset contains samples from the UK and Canada but represents humans in general as biological beings. Both datasets are incomplete (having missing values), so random forest imputation was performed to prevent information loss. The term *Survival Time* in this research always means how many months takes from patient initial diagnosis to their death.

The analysis uses the semi-parametric Cox Proportional Hazards model to observe whether a variable increases or decreases the risk of death or relapse for breast cancer patients. The response variables used are Survival Time (time from initial diagnosis to patient's death) and Relapse-Free Time (time from initial diagnosis to patient's first relapse), which are analyzed separately. This addresses the research gap from article [6] which used relapse-free time as a predictor instead of response variable. The Concordance Index is an evaluation metric for predictors in a survival model that determines whether the predictor significantly influences the response variable or just not more than a random event.

The calculation and algorithm execution were assisted using R Studio software with relevant libraries. The research procedure is as follows:

a. Retrieving time-to-event data and patient characteristics from the cBioPortal repository;
b. Performing data pre-processing (row deletion and Random Forest imputation);
c. Determining the survival time and relapse-free time variables as response variables and other variables that can be known at the time of initial diagnosis as predictor variables;
d. Testing the proportional hazards (PH) assumption for all predictors against both response variables separately using the Schoenfeld test;
e. Clustering the data using the K-Medoids method;
f. Comparing the PH assumption test results before and after clustering;
g. Creating Kaplan-Meier Curves for the clustered data and performing the log-rank test;
h. Building Cox proportional hazards models for each response variable;
i. Identifying significant predictors based on the Cox regression results;
j. Analyzing the hazard ratio of each significant predictor;
k. Evaluating the models using the Concordance Index, Likelihood Ratio test, and Wald test.

When the proportion of missing values in certain variable exceeded 10%, listwise deletion was applied to ensure consistency in model estimation. While this approach avoids additional assumptions regarding missing data mechanisms, it may introduce selection bias if the excluded observations differ systematically from those retained. Consequently, the analytical sample may not fully represent the underlying patient population, potentially limiting the generalizability of the findings. Tumor Stage was also excluded due to substantial missingness, which could compromise model stability if retained, despite its clinical relevance as a prognostic factor.

Regarding the missing data mechanism, no formal statistical assessment was conducted to distinguish between MCAR, MAR, or MNAR. The handling of missing values relied on the default behavior of the randomForest-based imputation procedure, which implicitly assumes that missingness is at least missing at random (MAR) given the observed covariates. If the missingness mechanism deviates from this assumption, particularly under an MNAR process, bias cannot be ruled out, and the generalizability of the results should therefore be interpreted with caution.

## 3. RESULT AND ANALYSIS

### 3.1 Pre-processing Data Result

METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset have samples that lack survival time and relapse-free time values. There are 529 of 2,509 samples (21.08%) on those two variables are missing values therefore they were removed and leaving 1980 samples. Furthermore, missing values were found in other variables as shown in table 1 which are expressed as percentages of missing values in each variable. The tumor stage variable was excluded from this study because it contains 36.27% missing values. Missing data in other variables were imputed using random forest with n-tree=1000. The final total of rows in the dataset are 1980 samples without missing value after being imputed with Random Forest Imputation. The Random Forest classifier achieved an OOB error of 0.186, indicating good generalization performance.

**Table 1.** Proportion of Missing Values for Each Variable

| Variable | Missing Value (%) | Variable | Missing Value (%) |
|---|---|---|---|
| Age at Diagnosis | 0 | Overall Survival (Months) | 0 |
| Cellularity | 3.34 | Overall Survival Status | 0 |
| ER Status | 0 | PR Status | 0.05 |
| Neoplasm Histologic Grade | 4.65 | Relapse Free Status (Months) | 0 |
| HER2 Status | 0.05 | Relapse Free Status | 0 |
| Primary Tumor Laterality | 5.94 | Tumor Size | 1.33 |
| Lymph nodes examined positive | 3.99 | Tumor Stage | 36.27 |

### 3.2 Survival Time as Dependent

#### 3.2.1 PH Assumption Before Clustered

Directly conducting the Schoenfeld test on imputed data resulted almost all variables not meeting the proportional hazards assumption, as shown in Table 2. The only variables that met the assumption were Cellularity, Primary Tumor Laterality, and Tumor Size. The global model also did not meet the assumption, so the overall features could not be modeled using cox regression.

**Table 2.** Schoenfeld Test Result

| Variable | P-value |
|---|---|
| Age at Diagnosis | < 2e-16 |
| Cellularity | 0.16956 |
| ER Status | < 2e-16 |
| Neoplasm Histologic Grade | 2.4e-11 |
| HER2 Status | 0.00014 |
| Primary Tumor Laterality | 0.76737 |
| Lymph nodes examined positive | 0.01735 |
| PR Status | 7.4e-13 |
| Tumor Size | 0.24729 |
| **Global** | **< 2e-16** |

Even after removed variables that does not met proportional hazard assumption, there still variable not meet the assumption i.e. Tumor Size as shown in table 3. P-value for the global test is very small, where the value exactly at the critical point of significance 0.05. Concordance index of 0.614 demonstrated discriminative ability on model.

**Table 3.** Schoenfeld Test Results After Removing Non-Significant Variables

| Variable | P-value |
|---|---|
| Cellularity | 0.061 |
| Primary Tumor Laterality | 0.752 |
| Tumor Size | 0.047 |
| Global | 0.050 |

#### 3.2.2 K-Medoids Clustering

Unsupervised machine learning was conducted as a way to meet the proportional hazards assumption by clustering the Age at Diagnosis variable and performing survival analysis for each cluster created. The elbow method considered the optimal number of clusters is 3 as in figure 2, which were labelled as Low-Age Cluster, Middle-Age Cluster, and High-Age Cluster. Number of observations in each cluster, respectively, from low to high cluster is 586, 706, and 688. Figure 1 shows that the middle-age cluster had longer maximum survival times (in months) than the low-age and high-age cluster. Notably, the clustering algorithm was only run on the Age at Diagnosis variable as one-dimensional clustering and y-axis at figure 1 is shown just for comparing survival time between cluster.

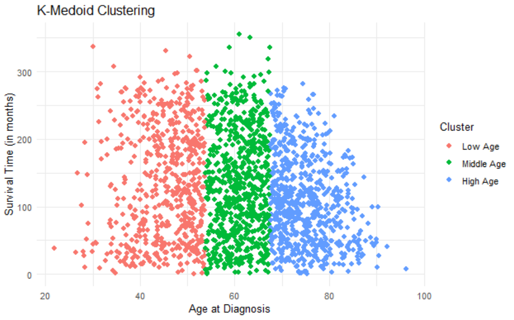

**Figure 1.** K-Medoids Clustering for Age at Diagnosis

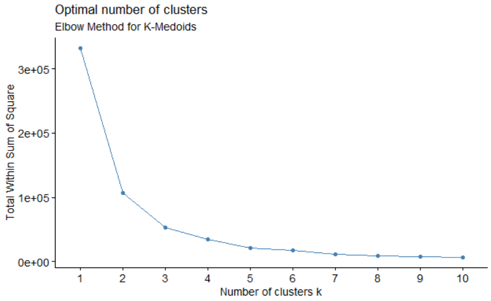

**Figure 2.** Elbow Method for Optimal Cluster

Table 4 provides information on class boundaries and average survival times for each cluster. The low age limit of 53.69 years was not chosen arbitrarily as in previous studies [3] but rather is the result of a more objective from euclidean distance calculation using the k-medoid clustering algorithm. The average survival time increases from the low-age cluster to the high-age cluster.

**Table 4.** Descriptive Statistics for Age at Diagnosis Clusters by Survival Time

| Cluster | Medoid | Min. Age | Max Age | Average Age | Max Survival Time (in months) | Average Survival Time (in month) |
|---|---|---|---|---|---|---|
| Low Age | 46.44 | 21.93 | 53.69 | 45.17 | 337.03 | 135.71 |
| Middle Age | 60.96 | 53.72 | 67.48 | 60.91 | 355.20 | 132.62 |
| High Age | 74.07 | 67.54 | 96.29 | 74.83 | 281.37 | 108.57 |

From a modeling perspective, violations of the proportional hazard assumption may arise when the effect of covariates varies implicitly with age. In a global Cox model, this induces time-varying or age-dependent regression effects, even when covariates are treated as time-invariant. By clustering patients using age as a one-dimensional variable, the population is partitioned into more homogeneous subgroups in which covariate effects are approximately constant over time. Within each age-based cluster, the Cox model can therefore be interpreted as a local approximation, where proportional hazards hold more closely than in the aggregated population.

Conceptually, age-based clustering serves a similar purpose to stratified Cox models or time-dependent covariates. Stratified Cox regression allows baseline hazards to vary across strata but does not estimate covariate effects within each stratum, while time-dependent Cox models explicitly relax the proportional hazards assumption at the cost of increased model complexity and reduced interpretability. In contrast, age-based clustering offers a compromise by maintaining standard Cox model interpretation while mitigating non-proportionality through population segmentation.

### 3.2.3 Kaplan-Meier Curve

Figure 3 shows Kaplan-Meier curve for each cluster created with death as the event. Those three clusters have significant differences in estimate the probability particularly after 50 months of survival time. Log-rank test conducted and gives p-value < 0.0001 means that all cluster significantly different in estimate the probability of breast cancer patient's survival time. Thus, clustering this data indeed important because patients with different age groups are estimated significantly different probability than another. At 300 months of survival time, high-age have not patient at risk anymore while low-age still have 6 and middle-age were 7.

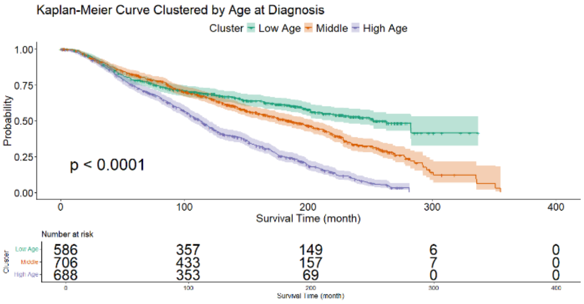

**Figure 3.** Kaplan-Meier Curves for Survival Analysis of Clustered Data

### 3.2.4 PH Assumption on Clustered Data

After data being clustered, Schoenfeld test re-conducted as detailed on Table 5 to compare the result after clustered with the result before clustered. Before data was clustered, only two variables (Cellularity and Primary Tumor Laterality) met the proportional hazard assumption, meanwhile after the clustering, the number of variables meeting the assumption increased to five for low-age and six for middle-age and high-age. Cellularity, HER2 Status, and Lymph nodes examined positive met the assumption in all three cluster. Age at Diagnosis only violates the assumption in the low-age cluster while Primary Tumor Laterality fails in high-age cluster and Tumor Size fails in middle-age cluster. Neoplasm Histologic Grade only satisfies the assumption in middle-age cluster and PR Status only satisfies it in high-age cluster.

**Table 5.** Schoenfeld Test Result for Clustered Data by Age

| Variable | P-value | | |
|---|---|---|---|
| | Low-Age | Middle-Age | High-Age |
| Age at Diagnosis | 0.00016 | 0.46354 | 0.4407 |
| Cellularity | 0.10919 | 0.58376 | 0.9902 |
| ER Status | 4.5e-10 | 0.00079 | 6.5e-05 |
| Neoplasm Histologic Grade | 1.3e-09 | 0.06890 | 0.0185 |
| HER2 Status | 0.06369 | 0.37967 | 0.7421 |
| Primary Tumor Laterality | 0.74281 | 0.38904 | 0.0393 |
| Lymph nodes examined positive | 0.15569 | 0.45290 | 0.8749 |
| PR Status | 9.3e-09 | 0.00712 | 0.0508 |
| Tumor Size | 0.46802 | 0.00144 | 0.9310 |
| Global | 2.0e-09 | 0.00774 | 0.0087 |
| Global (exclude non-significant predictor) | 0.157 | 0.373 | 0.72 |

This means that clustering data can increase the number of variables meeting the proportional hazard assumption but still interpretable since grouping patient ages is a common practice both academically and practically. After removing the non-significant variables and conducting the assumption test again, the real global p-value for the model was obtained for each three cluster, which each value far greater than the value of assumption test before the data was clustered.

### 3.2.5 Cox Regression for Low-Age Cluster

Based on cox regression modeling for Survival Time (time from initial diagnosis to death) as dependent variable, HER2 positive status, number of Lymph nodes that examined positive, and Tumor Size significantly affect how long breast cancer patients take from initial diagnosis to their death. Patients with positive HER2 Status had 1.8 times higher risk if compared to HER2 negative status. Number of Lymph nodes examined positive (HR>1) also positively associated with Survival Time hazard function. Similarly, higher Tumor Size (HR>1) corresponded to lower Survival Time. The full result of cox regression provided in table 6. The model demonstrated an acceptable discriminative ability with a concordance index of 0.691 (SE=0.017) which increased as compared to before data was clustered. Moreover, the likelihood ratio was 75.31 (df = 6, p < 0.001) and overall Wald test was 107.8 (df = 6, p < 0.001).

**Table 6.** Low-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| Cellularity.L | 1.165378 | 0.862-1.575 | 0.319814 |
| Cellularity.Q | 0.976131 | 0.762-1.250 | 0.848244 |
| HER2 Status (Positive) | 1.801023 | 1.311-2.474 | 0.000281 |
| Primary Tumor Laterality (Right) | 1.070994 | 0.827-1.388 | 0.603684 |
| Lymph nodes examined positive | 1.092547 | 1.064-1.121 | 2.86e-11 |
| Tumor Size | 1.009012 | 1.003-1.015 | 0.003209 |

### 3.2.6 Cox Regression for Middle-Age Cluster

For breast cancer patient with age from 53.72 to 67.48 years old when the initial diagnosis, HER2 positive status and number of Lymph nodes that examined positive significantly affect how long breast cancer's patients take from initial diagnosis to their death. Unlike low-age cluster, Tumor Size was not tested in this cluster because Tumor Size variable does not meet the proportional hazard assumption. Patients with positive HER2 Status had 1.45 times higher risk if compared to HER2 negative status. Number of Lymph nodes examined positive (HR>1) also positively associated with survival time hazard function. The full result of this cluster provided in table 7. The model demonstrated an acceptable discriminative ability with a concordance index of 0.629 (SE=0.016) which increase as compared to before data was clustered. Moreover, the likelihood ratio was 58.81 (df = 8, p<0.001) and overall Wald test was 79.87 (df = 8, p < 0.001).

**Table 7.** Middle-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| Age at Diagnosis | 1.0236 | 0.996-1.052 | 0.0924 |
| Cellularity.L | 0.8535 | 0.664-1.098 | 0.2174 |
| Cellularity.Q | 1.0377 | 0.856-1.257 | 0.7056 |
| Neoplasm Histologic Grade.L | 1.2711 | 0.950-1.702 | 0.1069 |
| Neoplasm Histologic Grade.Q | 1.1141 | 0.900-1.379 | 0.3208 |
| HER2 Status (Positive) | 1.4494 | 1.090-1.926 | 0.0106 |
| Primary Tumor Laterality (Right) | 1.0310 | 0.840-1.265 | 0.7705 |
| Lymph nodes examined positive | 1.0528 | 1.037-1.068 | 2.39e-12 |

### 3.2.7 Cox Regression for High-Age Cluster

For patient above 67.54 years old when the initial diagnosis, Age at Diagnosis, number of Lymph nodes that examined positive, and Tumor Size significantly affect how long breast cancer patients take from initial diagnosis to their death. Only at the high-age cluster HER2 positive status does not significantly increase the risk of survival time. Patients with higher age when diagnosis had 1.07 times higher risk to have shorter survival time. Number of Lymph nodes examined positive (HR>1) also positively associated with survival time hazard function in high-age cluster. Likewise, higher tumor size when diagnosis corresponded with higher mortality time of patient. The full result of this cluster provided in table 8. The model demonstrated an acceptable discriminative ability with a concordance index of 0.649 (SE=0.013) which increase as compared to before data was clustered. Moreover, the likelihood ratio was 136.4 (df = 7, p < 0.001) and overall Wald test was 159.2 (df = 7, p < 0.001).

**Table 8.** High-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| Age at Diagnosis | 1.0723 | 1.0548-1.090 | < 2e-16 |
| Cellularity.L | 0.9439 | 0.7647-1.165 | 0.5907 |
| Cellularity.Q | 1.1489 | 0.9722-1.358 | 0.1032 |
| HER2 Status (Positive) | 1.3012 | 0.9341-1.813 | 0.1196 |
| Lymph nodes examined positive | 1.0749 | 1.0507-1.100 | 4.91e-10 |
| PR Status (Positive) | 0.8468 | 0.7065-1.015 | 0.0718 |
| Tumor Size | 1.0091 | 1.0046-1.014 | 6.03e-05 |

## 3.3 Relapse-Free Time as Dependent

### 3.3.1 PH Assumption Before Clustered

Changing the dependent variable with Relapse-free Time (time from initial diagnosis to first relapse) and directly conducting the Schoenfeld test on imputed data resulted 4 variables fulfil the PH assumption as shown in Table 9. The only variables that met the assumption were Cellularity, Primary Tumor Laterality, Lymph nodes examined positive, and Tumor Size. But the global model if relapse-free set as dependent met the assumption far from critical point unlike in survival time. Concordance index of 0.642 demonstrated discriminative ability on model.

**Table 9.** Schoenfeld Test Results for Relapse-Free Time

| Variable | P-value |
|---|---|
| Age at Diagnosis | 0.00013 |
| Cellularity | 0.38987 |
| ER Status | < 2e-16 |
| Neoplasm Histologic Grade | 4.4e-08 |
| HER2 Status | 0.02218 |
| Primary Tumor Laterality | 0.80671 |
| Lymph nodes examined positive | 0.56219 |
| PR Status | 8.8e-13 |
| Tumor Size | 0.15649 |
| **Global** | **< 2e-16** |
| **Global (exclude non-significant predictor)** | **0.46** |

### 3.3.2 K-Medoids Clustering

Clustering performed at Age at Diagnosis variable, so there's no change in member of each cluster. Figure 1 shows comparably the Relapse-free Time each cluster, that the middle-age cluster had longer maximum relapse-free time (in months) than the low-age and high-age cluster.

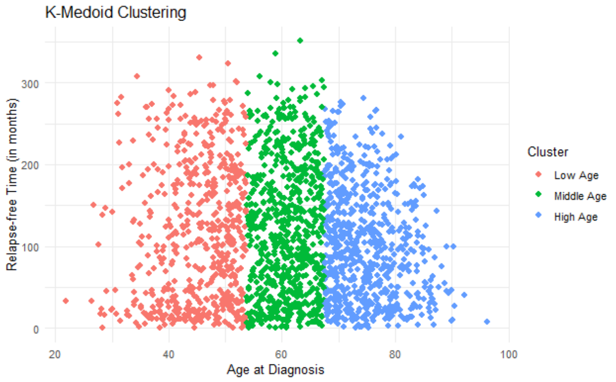

**Figure 4.** K-Medoids Clustering for Relapse-Free Time

Table 10 provides no different information with table 4 unless the Average Relapse-free Time (in month) column. Unlike the Average Survival Time that decreases as age cluster increases, Average Relapse-free Time here even have the highest value in the middle-age cluster with the second higher is low-age cluster. Max relapse-free time just likely max survival time that have highest value at middle-age cluster.

**Table 10.** Descriptive Statistics for Age at Diagnosis Clusters by Relapse-Free Time

| Cluster | Medoid | Min. Age | Max Age | Average Age | Max Relapse-free Time (in months) | Average Relapse-free Time (in month) |
|---|---|---|---|---|---|---|
| Low Age | 46.44 | 21.93 | 53.69 | 45,17 | 330.37 | 116.14 |
| Middle Age | 60.96 | 53.72 | 67.48 | 60,91 | 351.00 | 118.72 |
| High Age | 74.07 | 67.54 | 96.29 | 74,83 | 281.36 | 100.05 |

### 3.3.3 Kaplan-Meier Curve

Figure 5 shows Kaplan-Meier curve for relapse as event. Unlike in Survival Time section, the three clusters created did not show significant differences in estimated relapse probability. Log-rank test conducted and gives p-value 0,21, means these 3 clusters do not necessarily benefit from clustering because have relatively same estimated probability. This concludes that age group significantly different between cluster at death risk, but not in relapse risk of patient. However, a cluster-specific analysis still be conducted to assess for differences in the adherence to the proportional hazards (PH) assumption when the data are clustered. This stratification also facilitates intra-cluster comparison between the hazards associated with Survival Time and Relapse-Free Survival of each cluster.

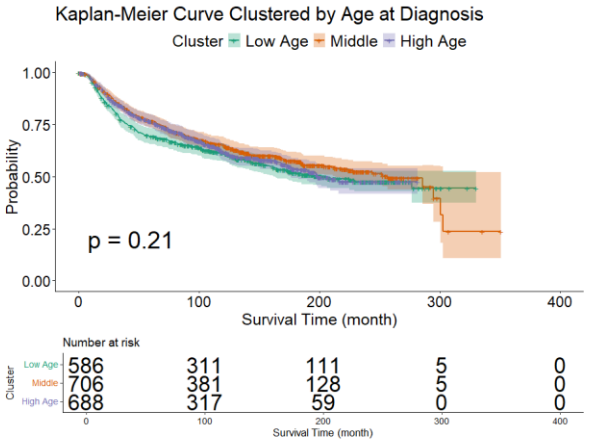

**Figure 5.** Kaplan-Meier Curves for Relapse-Free Time in Clustered Data

### 3.3.4 PH Assumption on Clustered Data

After data being clustered, Schoenfeld test re-conducted as detailed on Table 11 to compare the result after clustered with the result before clustered. Before data was clustered, only four variables met the proportional hazard assumption, meanwhile after the clustering, the number of variables meeting the assumption increased to 5 for middle-age, 6 for high-age, but remain 4 in low-age cluster. HER2 Status and Primary Tumor Laterality met the assumption in all three cluster. Age at Diagnosis and Cellularity only violates the assumption in the low-age cluster, while Lymph nodes examined positive fails in high-age cluster, and Tumor Size fails in middle-age cluster. Neoplasm Histologic Grade only satisfies the assumption in high-age cluster. In comparison with Survival Time section, Cellularity does not meet the assumption in low-age cluster at relapse-free time as dependent. Then in Neoplasm Histologic Grade, it changes from middle-age (Survival Time) to high-age (Relapse-free Time). PR Status does not meet even one cluster in Survival Time, but met at high-age at Relapse-free Time as dependent.

**Table 11.** Schoenfeld Test Results for Clustering Data by Age for Relapse-Free Time

| Variable | P-value | | |
|---|---|---|---|
| | Low-Age | Middle-Age | High-Age |
| Age at Diagnosis | 0.026 | 0.12955 | 0.4756 |
| Cellularity | 0.026 | 0.89512 | 0.4927 |
| ER Status | 2.4e-09 | 2.7e-06 | 0.0002 |
| Neoplasm Histologic Grade | 1.0e-06 | 0.00091 | 0.9474 |
| HER2 Status | 0.116 | 0.24051 | 0.8311 |
| Primary Tumor Laterality | 0.462 | 0.29331 | 0.4253 |
| Lymph nodes examined positive | 0.401 | 0.98500 | 0.0167 |
| PR Status | 1.2e-07 | 0.00012 | 0.0318 |
| Tumor Size | 0.179 | 0.00852 | 0.0755 |
| Global | 5.2e-08 | 3.7e-05 | 0.0142 |
| Global (exclude non-significant predictor) | 0.14 | 0.72 | 0.95 |

In Relapse-free Time as dependent also show that clustering data can increase the number of variables meeting the proportional hazard assumption. Even though low-age cluster remain at 4 variables met the assumption, but middle-age and high-age cluster increased to 5 and 6 respectively. After removing the non-significant variables and conducting the assumption test again, the real global p-value for the model was obtained for each three cluster, and be found that middle-age and high-age are increased in p-value meanwhile low-age decreased but still met the proportional hazard assumption.

### 3.3.5 Cox Regression for Low-Age Cluster

Based on cox regression modeling for Relapse-free Time as dependent variable, number of Lymph nodes that examined positive and Tumor Size significantly affect how long breast cancer patients take from initial diagnosis to relapse event (Relapse-Free Time variable). Number of Lymph nodes examined positive (HR>1) positively associated with Relapse-free Time hazard function. Similarly, higher Tumor Size (HR>1) corresponded to lower Relapse-free Time. The full result of hazard ratio and its p-value provided in table 12. The model demonstrated an acceptable discriminative ability with a concordance index of 0.649 (SE=0.017) which slightly increased as compared to before data was clustered. Moreover, the likelihood ratio was 48.29 (df = 4, p < 0.001) and overall Wald test was 68.86 (df = 4, p < 0.001).

**Table 12.** Low-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| HER2 Status (Positive) | 1.363076 | 0.9970-1.864 | 0.0522 |
| Primary Tumor Laterality (Right) | 1.124557 | 0.8812-1.435 | 0.3455 |
| Lymph nodes examined positive | 1.075211 | 1.0492-1.102 | 6.21e-09 |
| Tumor Size | 1.007174 | 1.0014-1.013 | 0.0154 |

### 3.3.6 Cox Regression for Middle-Age Cluster

For patient with age from 53.72 to 67.48 years old when initial diagnosis, HER2 positive status and number of Lymph nodes that examined positive significantly affect how long breast cancer's patients take from initial diagnosis to their relapse. Unlike low-age cluster, Tumor Size was not tested in this cluster because it did not meet the proportional hazard assumption. Patients with positive HER2 Status had 1.554 times higher risk if compared to HER2 negative status, higher than when Survival Time as dependent variable. Number of Lymph nodes examined positive also positively associated (HR>1) with Relapse-free Time hazard function. Only at the middle-age cluster HER2 positive status significantly increased the risk of relapse occurred. The full result of this cluster's hazard ratio and its p-value provided in table 13. The model demonstrated an acceptable discriminative ability with a concordance index of 0.61 (SE=0.017). Moreover, the likelihood ratio was 50.47 (df = 6, p < 0.001) and overall Wald test was 73.16 (df = 6, p < 0.001).

**Table 13.** Middle-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| Age at Diagnosis | 1.023951 | 0.9968-1.052 | 0.08476 |
| Cellularity.L | 0.883506 | 0.6887-1.133 | 0.32965 |
| Cellularity.Q | 1.035405 | 0.8549-1.254 | 0.72185 |
| HER2 Status (Positive) | 1.554553 | 1.1742-2.058 | 0.00206 |
| Primary Tumor Laterality (Right) | 1.022453 | 0.8332-1.255 | 0.83158 |
| Lymph nodes examined positive | 1.056256 | 1.0414-1.071 | 3.26e-14 |

### 3.3.7 Cox Regression for High-Age Cluster

For patient above 67.54 years old when initial diagnosis, Neoplasm Histologic Grade and Tumor Size significantly affect how long breast cancer patients take from initial diagnosis to their relapse. Patients with higher Neoplasm Histologic Grade had 2.019 times higher risk to have shorter Relapse-free Time. Likewise, higher tumor size when diagnosis corresponded with lower relapse-free time of patient (HR > 1). The full result of this cluster provided in table 14. The model demonstrated an acceptable discriminative ability with a concordance index of 0.626 (SE=0.019). Moreover, the likelihood ratio was 49.63 (df = 8, p < 0.001) and overall Wald test was 56.61 (df = 8, p < 0.001).

**Table 14.** High-Age Cluster Hazard Ratio

| Variable | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|
| Age at Diagnosis | 1.021985 | 0.9982-1046 | 0.0700 |
| Cellularity.L | 1.248094 | 0.8808-1769 | 0.2127 |
| Cellularity.Q | 0.916719 | 0.7080-1187 | 0.5095 |
| \`Neoplasm Histologic Grade\`.L | 2.019313 | 12737-3202 | 0.0028 |
| \`Neoplasm Histologic Grade\`.Q | 0.749665 | 0.5548-1013 | 0.0606 |
| HER2 Status (Positive) | 1.506188 | 0.9948-2281 | 0.0530 |
| Primary Tumor Laterality (Right) | 0.806894 | 0.6300-1033 | 0.0892 |
| Tumor Size | 1.016897 | 10111-1023 | 1.14e-08 |

### 3.4 Discussion and Limitation

This study identified that age-based clustering significantly improves model performance for survival time but shows less distinct differentiation for relapse-free time, particularly in older clusters. A potential statistical explanation for this discrepancy is the presence of competing risks. In the High-Age cluster, the mortality rate is naturally higher; patients may die from other causes or from the cancer itself before a relapse event can occur or be observed. In standard Cox proportional hazards models, death without relapse is typically treated as censored data, which may bias the risk estimation for relapse in populations with high mortality. Future studies could employ Competing Risk Regression (e.g., Fine-Gray models) to better isolate the specific risk of relapse by accounting for death as a competing event.

Furthermore, this study has several limitations. First, the exclusion of samples with missing survival times (listwise deletion) and the removal of the Tumor Stage variable due to high missingness (36%) may reduce the clinical generalizability of the findings. Second, the imputation of other predictors was performed assuming data were Missing At Random (MAR); deviations from this assumption could influence parameter estimates. Finally, the benefits of age clustering appeared most pronounced in the younger cohort (Low-Age), suggesting that biological heterogeneity in older patients might require more complex modeling than age stratification alone.

## 4. CONCLUSION

Clustering the METABRIC breast cancer patient dataset by Age at Diagnosis can improve the validity of Cox proportional hazards models by enhancing adherence to the proportional hazard assumption for both the Survival Time and Relapse-Free Time dependent variables. The discriminatory performance improved in all clusters for the Survival Time as dependent variable, but only in the low-age cluster for the Relapse-Free Time as dependent variable. This approach highlights heterogeneity in mortality risk across age groups, while relapse risk appears less sensitive to age-based partitioning. Overall, the findings demonstrate that age-based clustering offers a simple and interpretable strategy to regularize Cox models under proportional hazards violations, serving as a practical alternative to more complex modeling approaches such as stratified or time-dependent Cox regression.

Differences between survival time and relapse-free time suggest that these outcomes may be governed by distinct underlying risk structures. In particular, the benefits of age-based clustering were most pronounced in younger patients, while improvements were limited in middle-aged and older groups. This asymmetry indicates that the effectiveness of age clustering may not generalize uniformly across all populations and should be interpreted with caution. This study is limited by its reliance on a single dataset and by methodological constraints discussed earlier. Future research may extend this framework by comparing clustering-based approaches with alternative Cox model extensions and by exploring additional statistical structures, such as competing risks, to better capture the dynamics of relapse outcomes across age groups.

## 5. REFERENCES

[1] International Agency for Research on Cancer (IARC), "Global cancer observatory 2022," 2022. Accessed: Sep. 19, 2025. [Online]. Available: https://gco.iarc.fr/today/en/fact-sheets-cancers

[2] U. Adiga, S. Vasishta, A. J. Augustine, K. Farzia, E. Venkataravikanth, and L. Ravi, "Transforming breast cancer prediction: advanced machine learning models for accurate prediction and personalized care," 2025. Accessed: Dec. 21, 2025. [Online]. Available: https://lifescienceglobal.com/pms/index.php/ijsmr/article/view/10575

[3] Y. Li *et al.*, "Integrated prognostic model for young breast cancer patients: insights from SEER, METABRIC, and TCGA databases," *Clin Breast Cancer*, Jul. 2025, doi: 10.1016/j.clbc.2025.07.015.

[4] N. R. Pradana Ratnasari, "Comparative study of k-mean, k-medoid and hierarchical clustering using data of tuberculosis indicators in Indonesia," *Indonesian Journal of Life Sciences*, vol. 5, no. 2, pp. 9–20, Sep. 2023, doi: 10.54250/ijls.v5i02.181.

[5] H. Thottathyl and K. K. Pavan, "Differential evolution model for identification of most influenced gene in breast cancer data," *Ingenierie des Systemes d'Information*, vol. 27, no. 3, pp. 487–493, Jun. 2022, doi: 10.18280/isi.270316.

[6] Y. Gu, M. Wang, Y. Gong, S. Jiang, C. Li, and D. Zhang, "Unveiling breast cancer risk profiles: a comprehensive survival clustering analysis empowered by an online web application for personalized medicine," May 25, 2023. doi: 10.1101/2023.05.18.23290062.

[7] Z. Zhu, M. Hoag, S. Julien, and S. Cui, "Estimating mortality of insured advanced-age population with Cox regression model," 2002. Accessed: Oct. 15, 2025. [Online]. Available: https://www.bibsonomy.org/bibtex/043388ab9ec4eb8e48ff187755d72437

[8] S. Berestizhevsky and T. Kolosova, "The Cox hazard model for claims data," *Variance: Advancing the Science of Risk*, vol. 13, no. 2, pp. 265–278, Accessed: Oct. 20, 2025. [Online]. Available: https://www.yieldwise.com/Cox-Hazard-Model-Berestizhevsky-Kolosova%20Variance%20Journal.pdf

[9] P. R. Kaukuntla, "Advancing life insurance pricing accuracy through mortality forecasting: a time-series and survival analysis approach," *International Journal of Multidisciplinary Research and Growth Evaluation*, vol. 2, no. 1, pp. 729–734, 2021, doi: 10.54660/.ijmrge.2021.2.1.729-734.

[10] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[11] D. J. Stekhoven and P. Bühlmann, "Missforest-non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[12] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Information System*, vol. 101, Nov. 2021, doi: 10.1016/j.is.2021.101804.

[13] A. V. Ushakov and I. Vasilyev, "Near-optimal large-scale k-medoids clustering," *Information Science*, vol. 545, pp. 344–362, Feb. 2021, doi: 10.1016/j.ins.2020.08.121.

[14] A. Sobrinho Campolina Martins, L. Ramos de Araujo, and D. Rosana Ribeiro Penido, "K-medoids clustering applications for high-dimensionality multiphase probabilistic power flow," *International Journal of Electrical Power and Energy Systems*, vol. 157, Jun. 2024, doi: 10.1016/j.ijepes.2024.109861.

[15] R. Klar, N. Arvidsson, and D. Rudmark, "Towards a new last-mile delivery system: cost and energy-optimized robot and van allocation," *Transportation Research Part E: Logistics and Transportation Review*, vol. 204, Dec. 2025, doi: 10.1016/j.tre.2025.104392.

[16] D. Hartama, W. Wanayumini, and I. S. Damanik, "Pengelompokan algoritma k-means dan k-medoid berdasarkan lokasi daerah rawan bencana di Indonesia dengan optimasi elbow, DBI, dan silhouette," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, Sep. 2024, doi: 10.47065/bits.v6i2.5851.

[17] Wildani Eko Nugroho, S. Dwi Kurniawan, Y. Febrian Sabanise, and P. Prayoga, "Use of the k-medoids algorithm for food clustering using nutritional value and evaluation of the elbow method and the Davies-Bouldin index method," *Ultima InfoSys : Jurnal Ilmu Sistem Informasi*, vol. 16, no. 1, p. 33, Jun. 2025, doi: 10.31937/si.v16i1.4226.

[18] K. Markhaba, T. Aizhan, A. Karlygash, Z. Zheniskul, and K. Indira, "Identification and characterization of earthquake clusters from seismic historical data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, pp. 1594–1604, Dec. 2024, doi: 10.11591/ijeecs.v36.i3.pp1594-1604.

[19] E. T.Lee and J. Wenyu Wang, "Statistical Methods for Survival Data Analysis," 3rd ed., Hoboken: John Wiley & Sons, Inc., 2003. doi: 10.1002/0471458546.fmatter.

[20] Y. Farida, E. A. Maulida, L. N. Desinaini, W. D. Utami, and D. Yuliati, "Breast cancer survival analysis using Cox proportional hazard regression and Kaplan-Meier method," *Jurnal Teori dan Aplikasi Matematika*, vol. 5, no. 2, pp. 340–358, Oct. 2021, doi: 10.31764/jtam.v5i2.4653.

[21] M. S. Molydah S and D. Danardono, "An additive subdistribution hazards model for competing risks data," *Media Statistika*, vol. 16, no. 2, pp. 194–205, May 2024, doi: 10.14710/medstat.16.2.194-205.

[22] A. Alabdallah, M. Ohlsson, S. Pashami, and T. Rögnvaldsson, “The concordance index decomposition: a measure for a deeper understanding of survival prediction models,” *Artificial Intelligence in Medicine*, vol. 148, Feb. 2024, doi: 10.1016/j.artmed.2024.102781.

[23] T. Therneau and E. Atkinson, “Concordance,” 2024.

[24] E. Longato, M. Vettoretti, and B. Di Camillo, “A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models,” *Journal of Biomed Inform*, vol. 108, Aug. 2020, doi: 10.1016/j.jbi.2020.103496.

[25] Faisal Siddiq and Mohammad Dokhi, “Survival analisis durasi menganggur angkatan kerja disabilitas yang mengalami berhenti bekerja akibat pandemi Covid-19,” *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 2, pp. 326–340, Dec. 2022, doi: 10.21009/JSA.06217.

[26] M. S. Khan *et al.*, “Statistical non-significance, likelihood ratio, and the interpretation of clinical trial evidence: insights from heart failure randomized trials,” *Journal of Cardiac Failure*, vol. 30, no. 12, pp. 1629–1632, Dec. 2024, doi: 10.1016/j.cardfail.2024.07.026.

[27] K. D. Deane, L. Van Hoovels, V. E. Joy, N. Olschowka, and X. Bossuyt, “From autoantibody test results to decision making: incorporating likelihood ratios within medical practice,” *Autoimmunity Reviews,* vol.23, May 01, 2024. doi: 10.1016/j.autrev.2024.103537.

[28] A. Basu, A. Ghosh, A. Mandal, N. Martín, and L. Pardo, “A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator,” *Electron J Statist*, vol. 11, no. 2, pp. 2741–2772, 2017, doi: 10.1214/17-EJS1295.

[29] cBioPortal, “Breast cancer (METABRIC, Nature 2012 & Nat Commun 2016),” cBioPortal For Cancer Genomics. Accessed: Sep. 15, 2025. [Online]. Available: https://www.cbioportal.org/study/summary?id=brca_metabric

[30] Christina Curtis *et al.*, “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” *Nature*, vol. 486, pp. 346–352, Apr. 2012, doi: https://doi.org/10.1038/nature10983.

[31] B. Pereira *et al.*, “The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes,” *Nat Commun*, vol. 7, May 2016, doi: 10.1038/ncomms11479.