doi: 10.21608/esju.2024.314785.1040 *Research Article*
The Egyptian Statistical Journal (ESJ), 68 (2): 86-103

Homepage: https://esju.journals.ekb.eg/

The Egyptian Statistical Journal

Print ISSN 0542-1748– Online ISSN 2786-0086

Evaluating Fit of some Survival Analysis Models with Application and Simulation

Doaa A. Abdo[1, *], Ahmed R. EL- Saeed[2], Alaa A. Abdelmegaly[3]

Received 22 August 2024; revised 8 October 2024; accepted 9 October 2024

Keywords

AIC, BIC, GLM, GLMM, Cox proportional hazard model, Survival models.

Abstract

This paper conducts a comprehensive analysis of information criteria such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) applied on survival analysis models, including the Cox Proportional Hazard Model, Linear Mixed Model (LMM), and Generalized Linear Mixed Model (GLMM). The aim is to identify the model that fits the data in the best way based on these criteria. The paper proposes the utilization of various survival models, including the Cox Proportional Hazard Model, LMM, and GLMM to handle non- linear data which leads to accurate parameter estimates and comparing them using the proposed criteria. The primary objective is to estimate the coefficients of these models using breast cancer data consisting of (96) patients. The model's accuracy is assessed using two statistical criteria including AIC and BIC. The paper's findings demonstrate that, based on both AIC and BIC, the GLMM is the best fit for the application study with a value (120.4) for AIC and a value (179.1) for BIC. Also, the simulation study conducts that the best model fit at probabilities (0.2, 0.8) and sample size (50) is GLMM with (55.30) for AIC and (64.86) for BIC under exponential distribution, and the LMM under Weibull distribution with (61.61) for both AIC and BIC.

1. Introduction

In statistical analysis, the process of model selection holds significant importance as it seeks to identify the optimal model from a set of candidate models based on specific criteria and given data. This is crucial for any model-based inference. The selection of a mis-specified model not only results in theoretically different interpretations of the data but also leads to inappropriate conclusions in various applications, including biased parameter estimation, differential item functioning (DIF), and improper person-fit assessment (Taylor, 2005).

The challenge of choosing the most appropriate model for survival regression analysis imposes a balance between model fit and complexity by incorporating the likelihood, number of parameters, and sample size. The most widely used criteria for this concern are Akaike's (AIC) and Bayesian information criterion (BIC). AIC and BIC are the most used measures in health technology assessment to determine the best model that fits the data and should be used for the prediction for

✉ Corresponding author*: doaaashour@mans.edu.eg.
[1] Department of Applied Statistics and Insurance, Faculty of Commerce, Mansoura University, Mansoura, Egypt.
[2] Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia.
[3] Higher Institute of Advanced Management Sciences and Computers, Al-Beheira, Egypt.

long- treatment effects (Bütepage et al., 2022). Both AIC and BIC give quantitative measures to evaluate and compare different survival models, which enable the researchers from making right decisions (Tracy, 2024).

Both AIC and BIC are constructed based on the likelihood of a penalty, and the distinction between them lies in the penalty term. The penalty term is determined by the effective number of parameters in the model, serving as a measure of model complexity. The effective number of parameters is influenced by the prior distribution, which imposes additional constraints on the parameter space, leading to a reduction in the effective dimension. Although the number of parameters is directly derived from the likelihood, the incorporation of the prior distribution introduces complexities that are reflected in the effective number of parameters (van der Linden et al., 2010).

So, to enhance the development of model selection criteria within the Bayesian framework, various selection criteria like Akaike's (1974) information criterion (AIC) have been introduced by Congdon (2007) and Schwarz (1978). Bayesian Information Criterion (BIC) comes into play when employing fully Bayesian estimation methods, such as MCMC algorithms.

There are various studies that used AIC and BIC in survival analysis and regression analysis with application on medical data. For instance, (Abo EL Nasr et al., 2024) introduced AIC and BIC to compare some linear regression models involving GAM, Beta, GAM Beta, Ridge and Beta Ridge, with application on breast cancer data to handle multicollinearity issues. This study concluded that GAM has the best performance for AIC of this type of data, while the simulation study showed the superiority of Beta regression model of BIC and the Ridge regression model outperformed others based on AIC.

Abdo et al., (2024) introduced some survival models involving the excess hazard model and multilevel excess hazard model to estimate the excess hazard rather than the overall hazard with application on simulated data from software to get accurate estimation using two statistical criteria AIC and BIC. This study demonstrated superior performance for multilevel excess hazard models for estimating excess hazard under AIC and BIC.

Pluchart et al., (2023) conducted a comparison of seven comorbidity scores in relation to the four-month survival of 633 lung cancer patients. Their analysis concluded that the Elixhauser score exhibited the lowest AIC and BIC values, along with the highest C-statistics and Harrell’s C-statistics. These findings suggest that the Elixhauser score is the most effective predictive model for estimating four-month survival within the study cohort.

Bütepage et al., (2022) assessed the performance of AIC and BIC when comparing six standard parametric survival models to extrapolate survival data with different levels of right censoring. The results showed that at high levels of censoring (70% or more), neither AIC nor BIC can guide the choice of a survival model and should be used as weighted criteria.

Gallacher et al., (2021) used AIC and BIC criteria for comparing fit and estimates of restricted mean survival time (RMST) of eight parametric models including, (exponential, Weibull, gamma, log-normal, log–logistic- Gompertz, generalized gamma and generalized F) with application on simulated follow-up data. This study demonstrated that BIC is the best criterion for selecting the best model when the follow-up is more mature.

Lumley and Scott (2015) presented model selection criteria such as AIC and BIC in modeling large complex surveys and how two criteria can be modified to treat complex samples. They concluded that these criteria are a good fit for treating complex surveys.

Moreover, Miecznikowski et al., (2010) proposed a comparative survival analysis of breast cancer microarray studies to determine the most important prognostic genetic pathways using the Cox proportional hazard model and comparing pathways through AIC and BIC. This study used Cox proportional hazard regression to discover the most significant variables correlated with risk. The study emphasized that using AIC and BIC criteria outperformed other measures for determining the most significant variable impact on risk over the Cox proportional hazard model.

In general, AIC and BIC tend to overfit the true model asymptotically. However, the maximum likelihood estimates (MLE) associated with the model selected by AIC remain consistent and asymptotically normal.

This paper introduces model evaluation using AIC and BIC. The subsequent sections follow this outline: Section two introduces model specifications. The model fit criteria and the estimation methods for the used models are presented in sections three and four, respectively. In section five, numerical outcomes and concluding remarks are provided. Finally, the simulation study is proposed in section six.

2. Model Specification

Survival analysis is a field analyzing the occurrence time of an event of interest. It is used to calculate the survival rate of patients after some treatment. For example, this can be applied to analyze cancer patients after receiving chemotherapy. Survival analysis models are used in different branches, especially in medicine to study the effects of lifestyle variables on the disease life span and impute the effect of interventions. For a disease of breast cancer, it represents a disease determined by malignant cell growth in the mammary glands. Men and women affected by breast cancers. Most breast cancers form occur for female shortly before, during, or after menopause, with three-quarters of all cases being diagnosed after age 50, and the most diagnosed cancers causing women's death (Abdul Rahman et al., 2024). In this section, we will discuss several statistical survival models that are designed to address issues inherent in ordinary linear regression.

2.1 Cox Proportional Hazard Model

The Cox proportional hazards model is a widely used regression model in medical research for examining the relationship between patient survival time and one or more predictor variables. Cox's hazard model is considered a semi-parametric model, imparting robustness to Cox's method. Another advantage of employing Cox's method is its relative ease in incorporating time-dependent covariates. The cox's model takes the form as follows:

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots .. + \beta_p X_{ip}\} \quad (1)$$

where $\lambda_i(t)$ is the hazard function for the i^{th} subject, $\lambda_0(t)$ is the baseline hazard function and $\beta_1, \dots, \beta_p$ are parameters to be estimated.

The hazard ratio (HR) is used in multiple logistic regression analysis to represent the ratio of the observed to expected events between two independent comparison groups (Deo and Sundaram, 2021). The formula for calculating the HR is expressed as follows:

$$\mathrm{HR} = \frac{\sum \text{Observed Events in Group (1)in t} \,/\, \sum \text{Expected Events in Group (1)in t}}{\sum \text{Observed Events in Group (2)in t} \,/\, \sum \text{Expected Events in Group (2)in t}}$$

Alternatively, it can be expressed as:

$$\mathrm{HR} = \frac{\sum \mathrm{O_{Exp,t}}}{\sum \mathrm{E_{Exp,t}}} \Big/ \frac{\sum \mathrm{O_{Unexp,t}}}{\sum \mathrm{E_{Unexp,t}}}$$

Or, in the context of treated and control groups:

$$\mathrm{HR} = \frac{\sum \mathrm{O_{Treated,t}}}{\sum \mathrm{E_{Treated,t}}} \Big/ \frac{\sum \mathrm{O_{Control,t}}}{\sum \mathrm{E_{Control,t}}}$$

Here, O represents observed events, E represents expected events, and t denotes the time period.
Indeed, the Cox Proportional Hazards regression model comes with specific assumptions to ensure its appropriate use:

- Independence of survival times: This assumption necessitates that the survival times between distinct individuals in the sample are independent.
- Multiplicative relationship: The Cox model assumes that the predictors have a multiplicative effect on the hazard. This means that each predictor influences the hazard by a proportional factor, rather than through an additive contribution.
- Constant hazard ratio over time: The assumption of a constant hazard ratio over time implies that the proportional relationship between hazards remains consistent throughout the study duration.

2.2 Linear Mixed Models

In applied statistics, linear mixed models stand out as a versatile framework for modeling various types of data, encompassing clustered, longitudinal, and spatial data. Their significance lies not only in their inherent capabilities but also as a foundational platform for more intricate model classes. For instance, they serve as a precursor to the development of more complex models like GLMMs, as evidenced by works such as Mcculloch (2003). Additionally, they contribute to the foundation of models like nonlinear mixed models, as demonstrated in the work of Pinheiro and Bates (2000).
Linear mixed models extend linear regression models, and many of the methods used for selecting mixed models are adaptations of those developed for linear regression. However, key differences exist between the two. In linear regression, responses are assumed to be independent, whereas in linear mixed models, responses are often dependent. This dependence affects model selection by reducing the effective sample size. Additionally, linear mixed models include both regression parameters, which describe the mean structure, and variance parameters, which account for sources of variability and the dependence structure within the data.
In the linear mixed model represented by the equation:

$$\mathrm{Y} = \mathrm{X}\beta + \mathrm{Z}\zeta\,\mathrm{u} + \Delta\,\mathrm{e} \qquad (2)$$

where:

- Y is a vector of observed responses.

- X is a known $n \times p$ matrix of covariates.
- Z is a known $n \times s$ matrix.
- u and e are unknown observed independent n-vectors of random variables, where each element has a mean of zero and a variance represented by the identity matrix I_n.
- β is a p-vector of unknown regression parameters.
- ζ is an $s \times s$ matrix containing q_r distinct unknown parameters.
- Δ is an $n \times n$ matrix containing q_s distinct unknown parameters.

Let $\psi = \Gamma \Gamma^T$ and $\Sigma = \Delta \Delta^T$, so we can express:

$$E(Y) = X\beta \quad \& \quad Var(Y) = Z\psi Z^T + \Sigma \tag{3}$$

The generality of this notion allows the matrix roots Γ and Δ to be symmetric matrices obtained by taking the square roots of the eigenvalues in the spectral decomposition of ψ or Σ.

2.3 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) are commonly used to analyze correlated non-Gaussian data. These models extend linear mixed models or hierarchical linear models to handle non-continuous responses, such as binary outcomes or counts. GLMMs are also referred to as hierarchical or multilevel generalized linear mixed models. The terms "random coefficients models" or "random effects models" are often used interchangeably to describe both linear and generalized linear mixed models (Rabe-Hesketh & Skrondal, 2002).

GLMMs extend the generalized linear model by incorporating random effects into the linear predictor alongside the usual fixed effects (Breslow & Clayton, 1993; Stroup, 2012; Jiang, 2007). This allows for a flexible approach to analyzing grouped data, where differences between groups are modeled as random effects (Fitzmaurice et al., 2011). These models are particularly useful for capturing variability within and between groups, providing a more nuanced understanding of the data.

The predominant model employed for discrete outcomes is the GLMM. Consider a scenario where a specific test is administered repeatedly over time to a group of children. The outcome y_{ij} observed at the time (age) t_{ij} is binary, indicating pass or fail. In the context of subject-specific regression models, a logistic model could be proposed for this case. For instance, one could assume y_{ij} to follow a Bernoulli distribution with a success probability π_{ij} that satisfies:

$$logit(\pi_{ij}) = \log\left[\frac{\pi_{ij}}{1-\pi_{ij}}\right] = (\beta_1 + u_i) + \beta_2 t_{ij} \tag{4}$$

So, for each person we have a logistic model separately.

The model allows the availability of all individuals to differ based on their ability to pass the test. The random effects u_i follow a normal distribution with a mean of zero. Given the random effects u_i, it is assumed that the elements of y_i are independent, following the density functions of an exponential family form:

$$f_i(y_{ij}|u_i) = exp\left[\left(y_{ij}\eta_{ij} - a(\eta_{ij})\right) + \emptyset C(y_{ij}, \emptyset)\right] \tag{5}$$

where: A and C are functions, and $\emptyset$ is the over dispersion parameter. With mean $E(y_{ij}|u_i) = a'(\eta_{ij}) = \mu_{ij}(u_i)$ and variance $Var(y_{ij}|u_i) = \emptyset\, a''(\eta_{ij})$, and where, apart from a link function h, a linear regression model with parameters β and u_i is used for the mean, i.e.:

$$h(\mu_i(u_i)) = X_i\beta + Z_i u_i \tag{6}$$

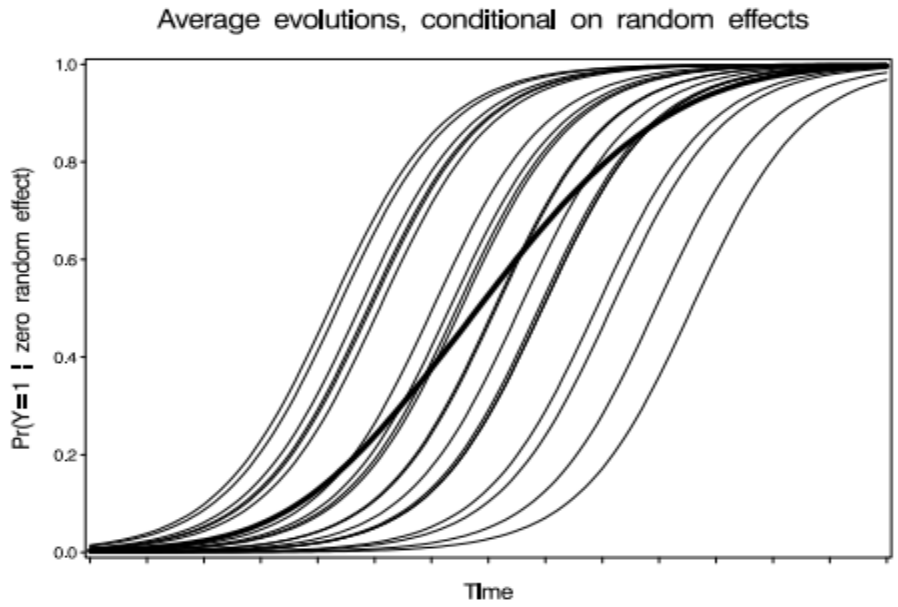

Figure 1: A random- intercepts curves of logistic model

Figure (1) presents a graphical representation of a random-intercepts logistic model. The thin lines in the figure correspond to the subject-specific logistic regression models for individual subjects, whereas the bold line represents the population-averaged trend (Verbeke & Molenberghs, 2013). This visualization highlights the variation among individual subjects while illustrating the overall trend across the population.

Briefly, the proposed survival models in this study have many applications not only in statistics but also in various branches like biology, medicine, education, and psychological science. Some of these applications are introduced in (Bolker et al., 2009), (Ko, 2017) and (Jiang et al., 2021).

3. Estimation Methodology

In this section, we will discuss estimation methods for the models previously presented in the previous section. It is noteworthy to mention here that the MLE is the approach used for estimation, given its statistical advantages that align with the structure of the parameters contained in those models.

3.1 Estimation of Cox Proportional Hazard Model

Parameter estimation in the Cox proportional hazards model is achieved by maximizing the partial likelihood, rather than the full likelihood. The partial likelihood can be expressed as:

$$L(\beta) = \prod_{Y_i} \frac{\exp(X_i\beta)}{\sum_{Y_j \geq Y_i} \exp(X_j\beta)} \tag{7}$$

The logarithm of the partial likelihood is expressed as:

$$l(\beta) = \log l(\beta) = \sum_{Y_i} \left\{ X_i\beta - \log\left[\sum_{Y_j \geq Y_i} \exp(X_j\beta) \right] \right\} \tag{8}$$

To obtain the MLEs of β by deriving the partial likelihood, the partial log-likelihood can be treated as an ordinary log-likelihood (Taylor, 2005).

3.2 Estimation of Linear Mixed Model

Maximum likelihood for this model is employed when linear models are not a good fit. Logistic regression is a specific instance of a GLM, and it is used for binary outcome data, where Y_i takes values of 0 or 1. The model is defined by the probability mass function given as (Shedden, 2010):

$$p(Y_i = 1|X_i = x) = \frac{\exp(\acute{\beta}X)}{1 + \exp(\acute{\beta}X)} = \frac{1}{1 + \exp(-\acute{\beta}X)} \quad (9)$$

which contains:

$$p(Y_i = 0|X_i = x) = 1 - p(Y_i = 1|X_i = x) = \frac{1}{1 + \exp(\acute{\beta}X)} \quad (10)$$

The log-likelihood for logistic regression:

$$L(\beta|y, X) = \log \prod \frac{exp(y_i.\acute{\beta}\, X_i)}{1 + \exp(\acute{\beta}X)} = \sum_{i,Y_i=1} \acute{\beta}X_i - \sum_i log(1 + \exp(\acute{\beta}X_i)). \quad (11)$$

This likelihood is utilized for the conditional distribution of Y given X by maximizing the above likelihood as a function of β. The gradient of the log likelihood function (the score function) is denoted as:

$$G(\beta|Y, X) = \frac{\partial}{\partial\beta} L(\beta|Y, X) = \sum_{i,Y_i=1} X_i - \sum_i \frac{exp(\acute{\beta}X_i)}{1 + \exp(\acute{\beta}X_i)} X_i \quad (12)$$

Therefore,

$$G(\beta|Y, X) = \sum_i \left(Y_i - \frac{exp(\acute{\beta}X_i)}{1 + \exp(\acute{\beta}X_i)}\right) X_i \quad (13)$$

3.3 Estimation of GLMMs

A linear mixed model takes the form

$$Y = X\beta + Zu + \varepsilon \quad (14)$$

where: $u \sim N(0, \sigma^2(D))$ and $\varepsilon \sim N(0, \sigma^2 I)$. Here D is an asymmetric and positive semidefinite matrix parameterized by a variance component vector θ, I is an $n \times n$ identity matrix, and σ^2 is the error variance. The conditional response of the dependent variable Y given β, u, θ and σ^2 is expressed as (Hariharan and Rogers ,2008):

$$Y|u, \beta, \theta, \sigma^2 \sim N(X\beta + Zu + \sigma^2 I_n) \quad (15)$$

The likelihood of Y given β, u, θ and σ^2 is defined as:

$$p(y|\beta, \theta, \sigma^2) = \int p(y|u, \beta, \theta, \sigma^2).p(u|\theta, \sigma^2)du \quad (16)$$

$$p(u|\theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \frac{1}{|D(\theta)|^{\frac{1}{2}}} exp\left\{\frac{-1}{2\sigma^2} u^T D^{-1} u\right\} \quad (17)$$

and

$$p(y|u, \beta, \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\left\{\frac{-1}{2\sigma^2}(Y - X\beta - Zu)^T(Y - X\beta - Zu)\right\}. \quad (18)$$

Let $\Lambda(\theta)$ is the lower triangular cholesky factor of $D(\theta)$ and $\Delta(\theta)$ be the inverse of $\Lambda(\theta)$. Then,

$$D(\theta)^{-1} = \Delta(\theta)^T \Delta(\theta) \quad (19)$$

Define

$$r^2(\beta,u,\theta) = u^T\Delta(\theta)^T\Delta(\theta) + (Y - X\beta - Zu)^T(Y - X\beta - Zu) \quad (20)$$

and suggest u^* is the value of u that satisfies:

$$\left.\frac{\partial r^2(\beta,u,\theta)}{\partial u}\right|_{u^*} = 0 \quad (21)$$

Then, for given β and θ, the likelihood function is given by:

$$p(y|\beta,\theta,\sigma^2) = (2\pi\sigma^2)^{\frac{-n}{2}} D(\theta)\Big|^{\frac{-1}{2}} exp\left\{\frac{-1}{2\sigma^2} r^2(\beta,u^*(\beta),\theta)\right\}.\frac{1}{|\Delta^T\Delta + Z^TZ|^{\frac{1}{2}}} \quad (22)$$

While the Cox proportional hazards model, linear mixed models, and generalized linear mixed models are commonly applied to longitudinal data analysis, Bayesian methods that rely on Markov Chain Monte Carlo (MCMC) techniques and non-informative priors are among the most popular estimation approaches for these models. In contrast, likelihood estimation for these models can be challenging to implement effectively.

Let $y_{(n)} = (y_1, y_2, \dots\dots, y_n)$ be the vector of data when n denotes the sample size. Consider the following general hierarchical model set-up:

Hierarchy 1: $y_{(n)}|X = x \sim h(y_{(n)}; X = x, \theta_1)$

Hierarchy 2: $X \sim g(x; \theta_2)$

We observe $y_{(n)}$ whereas X are unobserved. The parameters of interest are $\theta = (\theta_1, \theta_2)$.

To estimate the parameters θ and predict the unobserved states X. The likelihood function for this hierarchical model takes the form:

$$L(\theta, y_{(n)}) = \int h(y_{(n)}|x;\theta_1)\, g(x;\theta_2)dx \quad (23)$$

The difficulties related to applying this function for statistical inference are mainly computational:

1) The function of likelihood calculations includes high dimensional integration.
2) Assessing the location of the maximum using numerical search procedure is difficult because of estimated likelihood stochastic nature.
3) Difficulties in computing numerical computation of the second derivatives of the log likelihood function (Lele et al.,2010).

4. Model Selection Criteria

In the realm of model selection, we deal with a set of models $M_1, \dots, M_m$, where typically $m > 2$. These models may exhibit a "nested" structure with inclusions such as $M_1 \subset M_2 \subset \dots \subset M_m$, or they may not be nested. Instead of conducting multiple hypothesis tests between two models at a time - either rejecting one or accepting the other - it is more appropriate to have a criterion for selecting the most suitable model. To prevent overfitting, we opt for using the maximum log-likelihood rather than the maximum likelihood. Let ML_i represent the maximum likelihood for the i-th model, and $MLL_i = \log(ML_i)$, denote the maximum log-likelihood for the i-th model. Additionally, let d_i be the dimensions of the i-th model M_i. To avoid overfitting, various penalties have been introduced to subtract from MLL_i. The first was AIC or Akaike's information criterion which defined as:

$$AIC_i = MLL_i - d_i \quad (24)$$

Later, Schwarz (1978) introduced a different penalty, giving rise to the Bayes Information Criterion (BIC):

$$BIC = MLL_i - \frac{1}{2} d_i \log n \tag{25}$$

These criteria will be discussed in detail in the following section.

4.1 AIC Criterion

The Akaike Information Criterion (AIC) serves as a mathematical method for evaluating how well a model aligns with the data from which it was generated. In statistics, the AIC is employed to determine the optimal model for a given dataset. AIC depends on:

- The number of independent variables utilized in constructing the model.
- The MLE of the model.

In accordance with AIC, the most fitting model is the one that explains the highest amount of variation using the fewest independent variables possible (Bevans, 2020). The model with the minimum AIC value is chosen as the best-fitting model among competing models.

Akaike (1973, 1974, 1985, 1994) emphasized the importance of obtaining a robust model selection criterion based on Kullback-Leibler information (K-L). This involves estimating

$$E_y E_x \left[log \left(g\left(x|\hat{\theta}(y)\right)\right)\right] \tag{26}$$

where the inner part represents $E_f\left[log\left(g(x|\theta)\right)\right]$ with θ replaced by the MLE based on the assumed model g and data y. This double expectation is the objective of various model selection approaches based on $K-1$ information (e.g., AIC, AICc). Akaike established a connection between K-L information and likelihood theory, revealing that the maximum log-likelihood value was a biased estimate of $E_y E_x \left[\log\left(g\left(x|\hat{\theta}(y)\right)\right)\right]$. For large samples and good models, this bias is $\log\left(L\left(\hat{\theta}|data\right)\right) - K$. This result is equivalent to:

$$\log\left(L\left(\hat{\theta}|data\right)\right) - K = C - \hat{E}_\theta[I(f,\hat{g})] \tag{27}$$

where $\hat{g} = g\left(.|\hat{\theta}\right)$ and C is a constant.

This finding facilitates the integration of estimation methods (such as maximum likelihood or least squares) and model selection within a unified optimization framework. Akaike introduced an estimator for the expected relative Kullback-Leibler (K-L) information, which is corrected for asymptotic bias:

$$Relative\ \hat{E}(K-L) = \log\left(l\left(\hat{\theta}|data\right)\right) - K \tag{28}$$

where K is the term of asymptotic bias correction and multiplied this sample but profound result by -2, then the AIC is given by:

$$AIC = -2\log\left(l\left(\hat{\theta}|\ data\right)\right) + 2K \tag{29}$$

In the special case of least squares (LS) estimation with normally distributed errors, AIC can be expressed as:

$$AIC = n\log(\hat{\sigma}^2) + 2K \tag{30}$$

where $\hat{\sigma}^2 = \frac{\Sigma\left(\hat{\epsilon}_i\right)^2}{n}$.

4.2 Bayesian Information Criterion:

BIC, introduced by Schwarz in 1978, is an approximation method for Bayes computational factor. Due to the challenges associated with BF in model selection, BIC is often used as a substitute. Schwarz (1978) derived the BIC as:

$$\mathrm{BIC} = -2\log \mathrm{l}(\hat{\theta}|\mathrm{y}) + \mathrm{p}\log \mathrm{n} \tag{31}$$

where:

- $\hat{\theta}$ is the MLE of θ, that maximizes the likelihood function $l(\theta|y)$.
- p is the number of parameters in the model, i.e the dimension of $\theta, |\theta|$.
- n is the number of observations, i.e.$|y|$.

Typically, the BIC is calculated for each model, with the model exhibiting the lowest BIC value being selected. It is important to recognize that the term "BIC" is somewhat misleading, as it is not directly derived from information theory. Model comparison using BIC entails computing the BIC for each model and choosing the one with the lowest value.

Schwarz's BIC was originally justified under the assumption of independent and identically distributed (i.i.d) observations within linear models based on the regular exponential family likelihood. However, these constraints prompted further research. The original BIC has since been generalized to accommodate mixed effects models, where observations are correlated within subjects, as well as to other more complex models (Watanabe and Opper, 2010).

For two models, M_1 and M_2, BIC can be used to approximate the BF for model comparison:

$$\mathrm{BF} = \frac{\mathrm{p}(\mathrm{y}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{y}|\mathrm{M}_2)} = \exp\left\{\log\left[\frac{\mathrm{p}(\mathrm{y}|\mathrm{M}_1)}{\mathrm{p}(\mathrm{y}|\mathrm{M}_2)}\right]\right\} = \exp\{\log \mathrm{p}(\mathrm{y}|\mathrm{M}_1) - \log \mathrm{p}(\mathrm{y}|\mathrm{M}_2)\} \tag{32}$$

$$\therefore \mathrm{BF} \approx \exp\left\{-\frac{1}{2}\left(\mathrm{BIC}_1 - \mathrm{BIC}_2\right)\right\} = \exp\left\{-\frac{1}{2}\Delta\ \mathrm{BIC}\right\} \tag{33}$$

For mixed effects models containing both fixed and random effects, an improved BIC is employed. The improved BIC replaces the original BIC expression with the effective sample size:

$$\mathrm{BIC}_{\mathrm{n_e}} = -2\ \mathrm{logl}(\hat{\theta}|\mathrm{y}) + \mathrm{p}\ \log \mathrm{n_e} \tag{34}$$

where $\mathrm{n_e}$ is the effective sample size, defined as the magnitude of the correlation matrix (Berger et al. (2014)):

$$\mathrm{n_e} = |\mathrm{R}| = \frac{1}{1-\rho^2}(1-\rho+1-\rho) = \frac{2(1-\rho)}{1-\rho^2} = \frac{2}{1+\rho} \tag{35}$$

5. Numerical Analysis

This part introduced applying models on real data to estimate the models' parameters and selecting the best model based on this data, we relied on a sample of breast cancer data consisting of 95 patients. The sample obtained from an open source and downloaded from https://www.kaggle.com/search. This study included one dependent variable (event) represented the main concern and ten independent variables were Tumor size, Grade, Stage, Age, Sex, Cigarette, Packet per year, Type and Batch. All calculations were executed using R programming language version 4.2.2 and library survival in addition to library lme. We will outline these variables through figure (2) as follows:

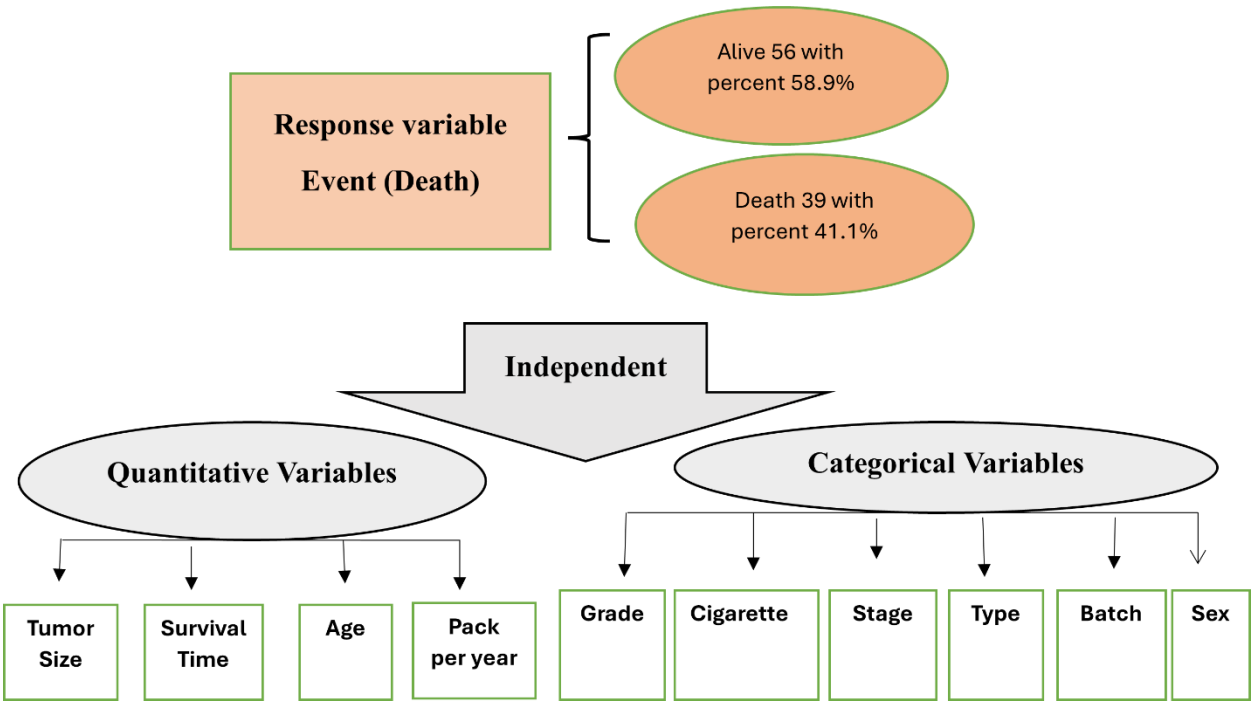

Figure 2: Study variables

Table (1) presents the number of observations, as well as the minimum, maximum, mean, and standard deviation for each variable in the breast cancer dataset. The results showed that the least standard deviation variable is the tumor size (Cm), which means that this variable has consistency and less differences between its observations then, age variable, after that the pack per year variable and finally the highest value of standard deviation is for survival time variable because there is variation between survival time of patients. The large maximum and minimum value was for the survival time variable.

Table 1: Descriptive statistics of quantitative variables

Variables	N	Minimum	Maximum	Mean	Std. Deviation
Tumor size (cm)	95	1.10	7.00	2.8547	1.36528
Survival time (days)	95	50.00	2532.00	1471.5158	681.64213
Age	95	48.00	88.00	66.5895	9.69810
Pack per year	95	0.00	105.00	29.2537	26.20578

Table (2) highlighted the parameter estimation of cox proportional hazard model and showed that the most significant variable was the survival time, and this variable was the least standard variable. This emphasizes that this variable the most influential variable on the occurrence of the death as, the survival rate increases the death rate decreases and so on.

Table 2: Cox proportional hazard model parameters estimation

Coefficient	Estimate	exp(coef)	se(coef)	Z value	Pr(>|z|)
Survival Time(days)	0.0009	1.0009	0.0003	3.6900	0.00022***
Tumor size	0.0303	1.0308	0.1180	0.2570	0.79731
Factor (Grade)2	-0.3289	0.7197	0.5333	-0.6170	0.53742
Factor (Grade)3	-0.4974	0.6081	0.5168	-0.9620	0.33585
Factor (Stage)2	-0.2701	0.7633	0.3668	-0.7360	0.46147
Factor (Stage)3	0.2717	1.3122	0.3462	0.7850	0.43245
Factor (Stage)4	-0.0771	0.9258	0.4809	-0.1600	0.87270
Factor (Stage)5	0.2510	1.2853	0.7920	0.3170	0.75134

Coefficient	Estimate	exp(coef)	se(coef)	Z value	Pr(>|z|)
Factor (Stage)6	-0.9320	0.3938	0.6909	-1.3490	0.17732
Factor (Stage)7	-0.0428	0.9581	0.4458	-0.0960	0.92346
Factor (Stage)8	-1.1077	0.3303	1.1415	-0.9700	0.33186
Factor (Stage)9	-0.2670	0.7657	0.7634	-0.3500	0.72651
Age	-0.0214	0.9789	0.0142	-1.5030	0.13282
Factor (Sex)1	-0.3807	0.6834	0.2853	-1.3340	0.18214
Factor (Cigarette)1	-0.4579	0.6326	0.3768	-1.2150	0.22433
Factor (Cigarette)2	-0.4727	0.6234	0.5253	-0.9000	0.36825
Pack per year	0.0031	1.0031	0.0076	0.4070	0.68379
Factor (Type Adjuvant)1	0.5936	1.8105	0.8754	0.6780	0.49770
Factor (Type Adjuvant)2	0.5493	1.7320	0.9014	0.6090	0.54229
Factor (Type Adjuvant)3	-0.2016	0.8175	1.0668	-0.1890	0.85014
Factor (Type Adjuvant)4	0.4372	1.5483	1.4003	0.3120	0.75491
Factor (batch)2	0.1623	1.1762	0.4050	0.4010	0.68869
Factor (batch)3	0.2611	1.2983	0.3409	0.7660	0.44374

Table (3) introduced the linear mixed model coefficients estimation, this table showed that the survival time was the most influential variable on the death event with less standard error, this proves on increasing the risk of the death related to the length of disease infection. Also, survival time, pack per year and tumor size variables have a negative relationship with death rate.

Table 3: Linear Mixed Model Parameters estimation

Coefficient	Estimate	Std Error	Z value	$Pr(>\|Z\|)$
Intercept	15.5700	4272	0.004	0.997054
Survival Time (days)	**-0.0025**	**0.0007**	**-3.464**	**0.0005*****
Tumor Size	-0.0183	0.3271	-0.056	0.955368
Factor Grade 2	0.2637	1.5140	0.174	0.861741
Factor Grade 3	0.65584	1.3980	0.469	0.639059
Factor Stage 2	0.5168	1.0640	0.486	0.627168
Factor Stage 3	-1.1560	1.0560	-1.094	0.273879
Factor Stage 4	-0.3892	1.4450	-0.269	0.787705
Factor Stage 5	-0.3409	2.0250	-0.168	0.866320
Factor Stage 6	3.2450	1.8460	1.758	0.078766
Factor Stage 7	0.1896	1.1990	0.158	0.874308
Factor Stage 8	19.3500	6523	0.003	0.997633
Factor Stage 9	17.5300	3281	0.005	0.995738
Age	0.0737	0.0412	1.789	0.073571
Sex	1.2340	0.7967	1.550	0.121089
Factor Cigarette 1	2.2150	1.1940	1.855	0.063664
Factor Cigarette 2	2.0580	1.4990	1.372	0.169935
Pack per year	-0.0232	0.0187	-1.232	0.217835
Factor (Type Adjuvant)1	-19.1300	4272	-0.004	0.996428
Factor (Type Adjuvant)2	-18.5000	4272	-0.004	0.996544
Factor (Type Adjuvant)3	0.4510	5155	0.000	0.999930
Factor (Type Adjuvant)4	-1.5290	7797	0.000	0.999844
Factor (batch)2	-0.7632	1.2630	-0.604	0.545663
Factor (batch)3	-0.7698	1.0200	-0.755	0.450450

Table (4) proposed the estimation of GLMM parameters. From this table we found that the most significant variables on the death risk were the survival time and the Stage variable especially stage 6 from breast cancer stages, was the most dangerous stage which raised the death risk. From the findings of GLMM estimates, we found that when there are stages and clusters in proposed data

of breast cancer, the GLMM determined the most dangerous stage of disease and how it influence in increasing the risk of death.

Table 4: GLMM parameters estimation

Coefficients	Estimate	Std. Error	Z value	$Pr(>\|Z\|)$
Intercept	15.410	11690	0.001	0.998948
Survival Time(days)	**-0.002**	**0.001**	**-3.690**	**0.0002*** **
Tumor size	0.013	0.314	0.041	0.967599
Factor Grade 2	0.414	1.502	0.276	0.782777
Factor Grade 3	0.851	1.369	0.621	0.534534
Factor Stage 2	0.703	0.960	0.733	0.463605
Factor Stage 3	-1.058	0.986	-1.073	0.283264
Factor Stage 4	-0.254	1.411	-0.180	0.857408
Factor Stage 5	0.004	2.055	0.002	0.998514
Factor Stage 6	**3.479**	**1.749**	**1.990**	**0.046615***
Factor Stage 7	0.185	1.093	0.169	0.865754
Factor Stage 8	26.250	342800	0.000	0.999939
Factor Stage 9	26.040	241000	0.000	0.999914
Age	0.074	0.041	1.802	0.071492
Sex	1.179	0.774	1.523	0.127671
Factor Cigarette 1	2.275	1.200	1.896	0.058003
Factor Cigarette 2	2.000	1.484	1.348	0.177507
Pack per year	-0.025	0.019	-1.332	0.182707
Factor (Type Adjuvant)1	20.000	11690	-0.002	0.998635
Factor (Type Adjuvant)2	-19.400	11690	-0.002	0.998676
Factor (Type Adjuvant)3	8.047	348600	0.000	0.999982
Factor (Type Adjuvant)4	4.793	411000	0.000	0.999991

6. Simulation Study

In this section, we proposed the selection model criteria including AIC and BIC criteria respectively for the previously proposed models. These criteria have been done under exponential and Weibull distributions with various four levels of sample size $n = (50, 250, 500, 800)$, and three different values of probability $p = (0.4, 0.5, 0.8)$ to show the differences between the criteria when increasing the sample size and the risk probability. We depended on small and large samples with small and large probabilities to generalize these results on the proposed models.

Table (5) and Table (6) viewed the calculations of both AIC and BIC criteria, respectively, for each model with two distributions (Exponential and Weibull) at three levels of $p = (0.4, 0.5, 0.8)$ and four different sample size $n = (50, 250, 500, 800)$. From these calculations, we found for both AIC and BIC the following findings:-

- At probability (0.2) and (0.5) for all sample sizes the value of criteria increased as the probability increased for all models
- After that at probability (0.8) the value decreased for all models' criteria with an increase in sample size to 800 units.

Hence, we can conclude that when the sample size increased, the number of observations rose and hence the differences decreased, this raised the accuracy of models through selection criteria.

Table 5: Simulation study: AIC values

Criteria		**AIC**					
Distribution		**Exp**			**Weibull**		
Model		**COX**	**LMM**	**GLMM**	**COX**	**LMM**	**GLMM**
$n = 50$	$p = 0.2$	305.39	56.09	55.30	301.04	53.96	55.76
	$p = 0.5$	307.71	76.51	74.76	301.27	72.77	74.53
	$p = 0.8$	307.63	56.09	55.30	302.22	53.96	55.76
$n = 250$	$p = 0.2$	2277.44	257.75	256.02	2272.12	255.29	257.15
	$p = 0.5$	2280.03	355.16	352.20	2272.62	350.57	352.36
	$p = 0.8$	2280.17	257.75	256.02	2273.46	255.29	257.15
$n = 500$	$p = 0.2$	5232.24	509.60	507.61	5226.72	504.60	506.40
	$p = 0.5$	5234.82	702.17	699.03	5227.21	697.15	698.95
	$p = 0.8$	5235.28	509.60	507.61	5228.06	504.60	506.40
$n = 800$	$p = 0.2$	9113.73	808.64	806.58	9108.02	805.18	807.03
	$p = 0.5$	9116.06	1118.12	1114.90	9108.43	1112.98	1114.79
	$p = 0.8$	9116.72	808.64	806.58	9109.32	805.18	807.03

Table 6: Simulation study: BIC values

Criteria		**BIC**					
Distribution		**Exp**			**Weibull**		
Model		**COX**	**LMM**	**GLMM**	**COX**	**LMM**	**GLMM**
$n = 50$	$p = 0.2$	317.22	69.83	64.86	306.77	61.61	65.32
	$p = 0.5$	322.18	92.89	84.32	307.01	80.42	84.09
	$p = 0.8$	319.46	69.83	64.86	307.96	61.61	65.32
$n = 250$	$p = 0.2$	2301.60	285.43	273.63	2282.68	269.37	274.76
	$p = 0.5$	2308.19	386.85	369.80	2283.19	364.66	369.97
	$p = 0.8$	2304.32	285.43	273.63	2284.02	269.37	274.76
$n = 500$	$p = 0.2$	5262.21	543.78	528.68	5239.37	521.46	527.47
	$p = 0.5$	5268.54	740.10	720.11	5239.85	714.00	720.02
	$p = 0.8$	5265.25	543.78	528.68	5240.70	521.46	527.47
$n = 800$	$p = 0.2$	9147.53	847.12	830.00	9122.08	823.92	830.46
	$p = 0.5$	9153.54	1160.28	1138.32	9122.48	1131.71	1138.22
	$p = 0.8$	9150.52	847.12	830.00	9123.37	823.92	830.46

Table (7) proposed the summary information criterions for each model AIC and BIC, for exponential distribution the best model fit was the GLMM at $n = 50$ with $p = 0.2, p = 0.8$, on the other hand when the data followed the weibull distributions the GLMM was the best model for both probability level 0.2 and 0.8 at sample size 50. for both criterions the best model fit was the generalized linear mixed model because it has the least value for all.

Table 7: Summary of Model selection criterions

Model	**AIC**	**BIC**
COX	321.3006	357.8989
LMM	121.7251	183.0181
GLMM	120.4	179.1

Table (8) introduced the residual deviance, minimum, the first and third quartile, median and maximum value for each model, the less residual deviance value was for the LMM with minimum value -3.3125 and maximum value 2.6241, which proved that this model treats the differences between the observations.

Table 8: Dispersion parameter of binomial family taken deviance residuals

Model	Residual deviance	Min.	Q1	Median	Q3	Max.
COX	379.04	-0.5062	-0.1541	-0.1045	-0.0692	3.8275
LMM	73.725	-2.2896	-0.6206	-0.2430	0.5709	1.8305
GLMM	3.996e-16	-3.3125	-0.4820	-0.1701	0.4828	2.6241

Figure (3) outlined the hazard function which presented the failure rate and gives the probability of event (occurrence of death) for breast cancer patients, the hazard function shown above is an example of a monotone increasing hazard between ages from (30 to 69). At age 69 the probability of hazard rate is 0.02, while at age 40 the hazard probability represented 0.01, this proved that this function is an increasing function, and the risk of death raised with time.

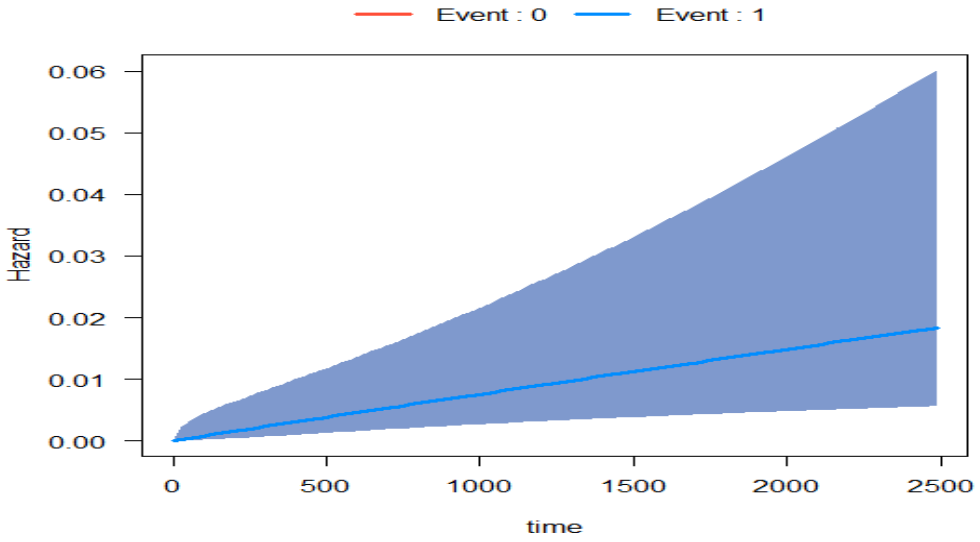

Figure 3: Hazard function of event and non- event for breast cancer data

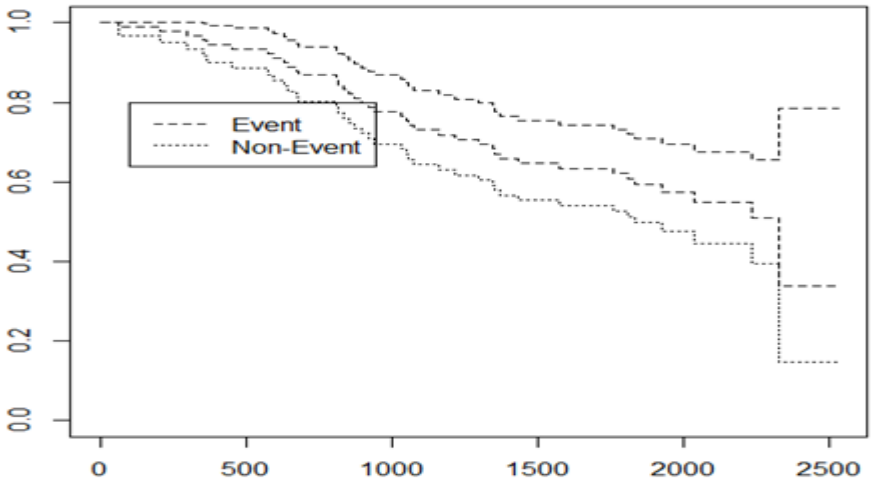

Figure 4: Survival function for breast cancer data

Figure (4) presents the survival function which represents the probability that a patient of interest will survive past a certain time. The previous figure displays survival function for live and death patients of breast cancer, and viewed that at age 60 or more the survival rate was 0.2, while at age 30 it was 0.8, this function decreased as time increased, which emphasizes on the risk of death increased with time.

Figures (5) and (6) illustrate the AIC and BIC curves at different sample sizes and various levels of probability with 10000 iterations. These curves showed that these values varied from increasing to decreasing and so on with the variation of sample size and probability.

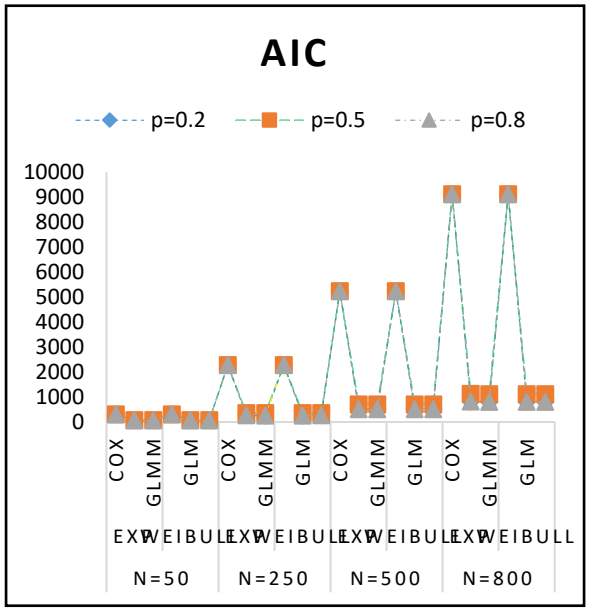

Figure 5: The AIC for estimated models

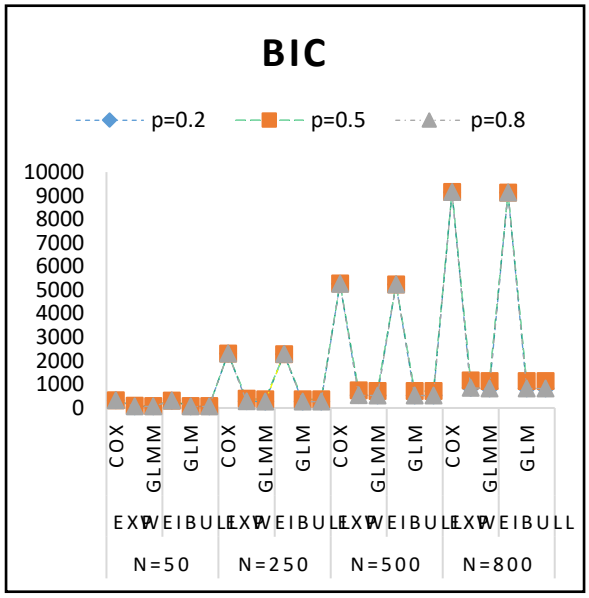

Figure 6: The BIC for estimated models

7. Conclusion

This research introduced several models, namely the Cox proportional hazard model, linear mixed model, and generalized linear mixed model, and applied them to real breast cancer data. The models' parameters were estimated using maximum likelihood. Additionally, a simulation study was conducted to generate AIC and BIC criteria across different sample sizes (50,250,500,800) and various values of $p = (0.2,0.4,0.8$. These models were applied to a sample consisting of 95 patients with ten explanatory variables and one binary response variable representing the occurrence of death or non- occurrence of death.

The results highlighted the most significant independent variables affecting the increased risk of death were the survival time and the breast cancer stage 6. Akaike´s information criteria (AIC) and Bayesian information criteria (BIC) were used to compare between these models. The comparison findings demonstrated that the best-fitting model was the generalized linear mixed model for both AIC and BIC. The simulation revealed that the GLMM for the exponential distribution and the LLM for the Weibull distribution were the best-fitting models. For future studies we will seek to use multilevel models under longitudinal data and panel data and apply these models in the

presence of some regression problems like multicollinearity. Also using various criteria to determine the best model fit like $\Delta\, AIC$ and ΔBIC rather than AIC and BIC.

Declaration of interests

The authors declare that they have no conflict of interest.

References

Abdo, D. A., Abd-Elmegaly, A. A., and Nasr, A. E. (2024). Survival models including excess hazard model and multilevel excess hazard model with application. *The Egyptian Statistical Journal*, 68(1): 78-90.

Abdul Rahman, H., Zaim, S. N. N., Suhaimei, U. S., and Jamain, A. A. (2024). Prognostic Factors Associated with Breast Cancer-Specific Survival from 1995 to 2022: A Systematic Review and Meta-Analysis of 1,386,663 Cases from 30 Countries. *Diseases*, 12(6), 111.

Abo El Nasr, M. M., Abdelmegaly, A. A., and Abdo, D. A. (2024). Performance evaluation of different regression models: application in a breast cancer patient data. *Scientific Reports*, 14(1): 12986.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6): 716-723.

Bevans, R. (2020). An introduction to the Akaike information criterion. (Accessed 4 May 2021). Downloaded from: https://www. scribbr. com/statistics/akaike-information-criterion/.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3): 127-135.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421): 9-25.

Bütepage, G., Vitor, C., & Carlqvist, P. (2022). MSR71 Performance of AIC and BIC for the extrapolation of survival data with different levels of censoring. *Value in Health*, 25(12): S363.

Congdon, P. (2007). Bayesian Statistical Modelling. John Wiley & Sons. Creative Commons Attribution Share Alike, 3: 35.

Deo, S. V., Deo, V., and Sundaram, V. (2021). Survival analysis- part 2: Cox proportional hazards model. *Indian Journal of Thoracic and Cardiovascular Surgery*, 37: 229-233.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). Fixed effects versus random effects models. Applied Longitudinal Analysis. Wiley, New Jersey.

Gallacher, D., Kimani, P., and Stallard, N. (2021). Extrapolating parametric survival models in health technology assessment: a simulation study. *Medical Decision Making*, 41(1): 37-50.

Hariharan, S., & Rogers, J. H. (2008). Estimation Procedures for Hierarchical Linear Models. In Multilevel Modeling of Educational Data (Eds. A. A. Connell and D. B. McCoach). Information Age Publishing, Inc.

Jiang, J., and Nguyen, T. (2007). Linear and Generalized Linear Mixed Models and their Applications 1. New York: Springer.

Jiang, J., and Nguyen, T. (2021). Linear Mixed Models: Part I. In: Linear and Generalized Linear Mixed Models and Their Applications. Springer Series in Statistics. Springer, New York, NY. doi.org/10.1007/978-1-0716-1282-8_1

Ko, J. (2017). Solving the Cox proportional hazards model and its applications (Doctoral dissertation, Master's thesis. EECS Department, University of California, Berkeley).

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492): 1617-1625.

Lumley, T., and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1): 1-18.

McCulloch, C. E. (2003). Generalized linear mixed models. Ims.

Miecznikowski, J. C., Wang, D., Liu, S., Sucheston, L., and Gold, D. (2010). Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC Cancer*, 10: 1-7.

Pinheiro, J. C., and Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus, Springer New York, NY, 3-56.

Pluchart, H., Bailly, S., Chanoine, S., Moro-Sibilot, D., Bedouch, P., and Toffart, A. C. (2023). Comparison of seven comorbidity scores on four-month survival of lung cancer patients. BMC *Medical Research Methodology*, 23(1): 256.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1): 1-21.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461-464.

Shedden, K. (2010). Generalized Linear Models. Department of Statistics, University of Michigan.

Stroup, W. W. (2012). Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. CRC press.

Taylor, J. (2005). Statistics 203: Introduction to Regression and Analysis of Variance—Model Selection: General Techniques.

Tracy, S. J. (2024). Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact. John Wiley & Sons.

van der Linden, W. J., Klein Entink, R. H., and Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34(5): 327-347.

Verbeke, G., and Molenberghs, G. (2013). The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, 14(3): 477-490.

Watanabe, S., and Opper, M. (2010). Asymptotic equivalence of Bayes cross-validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12).