Investigating accuracy of biomarker involving a parametric approach of proportional hazard skewed normal model

Ahmad Faiz Mohd Azhar[1*], and *Adina Najwa* Kamarudin[1]

[1]Universiti Teknologi Malaysia, Department of Mathematics, Faculty of Science, 81310 Skudai, Johor, Malaysia

Abstract. Time-dependent receiver operating characteristic (ROC) curve is useful to measure the accuracy performance over time. In this paper, we have shown how to determine the accuracy trend using proportional hazard model with continuous skewed normal biomarker and skewed normal time-to-event. Bayesian inference and adaptive multivariate integration over hypercubes are used respectively for parameter estimation and solving the sensitivity and specificity of the time-dependent ROC. The simulation study and application on real data suggest that it is possible to predict the accuracy measurement over time by changing the estimated association parameter between the biomarker and time-to-event data. In addition, studies on the impact of sample size on the ROC curve shows an advantage of this parametric method over conventional nonparametric.

1 Introduction

In recent years, we have observed a noteworthy development in time-dependent receiver operating characteristic (ROC) analysis. Studies relating to the time-dependent ROC curve and its area under the curve (AUC) have proven to benefit researchers, as both metrics can be used as tools to distinguish key variables that can effectively separate diseased or nondiseased populations. To recall, we can check the accuracy of a test with a binary outcome by computing sensitivity and specificity. In survival studies, sensitivity measures the probability of correctly classifying individuals with the disease into a diseased population. Meanwhile, specificity measures the probability of correctly classifying individuals without the disease into a nondiseased population. A specific biomarker value that can maximize both sensitivity and specificity is commonly used as a threshold to separate individuals into either population. However, since individuals' health status can change over time, extra consideration is required. As demonstrated by Etzioni et al. [1] using the binormal model to directly model the ROC curve might help solve issues related to the patient's status change over time. To use binormal model, each diseased and nondiseased population is first described through a normal distribution. Etzioni et al. [1] modified the binormal model by allowing a linear growth on the location for the case (diseased) population before directly

*Corresponding author: ahmad97@graduate.utm.my

generating the time-dependent ROC. However, using the binormal model might pose challenges because the ROC curve might produce an S-shape curve that can cross the random chance line (diagonal line), thus becoming an invalid ROC curve.

Heagerty et al. [2] refined the time-dependent ROC methodology through survival model and conditional distribution to simplify the procedure to generate a time-dependent ROC. Despite its ease of implementation, using the Kaplan-Meier (KM) to estimate the survival rate did not guarantee a monotonic time-dependent ROC curve. Using the nonparametric KM method produces an ROC curve resembling a step function, which brings complication when finding a unique threshold that can maximise both sensitivity and specificity rates. As exemplified by Unal [3], an additional method that uses the Youden's J index will need to be considered when finding the optimum point from a nonmonotone time-dependent ROC curve. Applying kernel smoothing techniques such as demonstrated by Heagerty and Zheng, Martinez-Camblor and Pardo-Fernandez, Beyenne and El Ghouch, and Diaz-Coto et al. [4–7] managed to solve issues related to monotonicity. Nonetheless, these methods require substantial sample sizes and proper kernel bandwidth selection to yield a robust ROC curve. This is because underestimation of the AUC might occur when using kernel bandwidth that might oversmooth the ROC curve. The AUC is defined as the probability of a randomly selected pair of subjects, with one diseased and one non-diseased, where the marker value of the diseased subject is greater than that of the non-diseased subject. A higher AUC indicates a better accuracy performance for the biomarker.

Implementing semiparametric techniques such as Cox regression facilitates solving issues around prospective studies, thus making it desirable to be used by researchers aiming to assess their model's predictive capabilities. The production of time-dependent ROC curve under the Cox regression method usually employs the definition of incident sensitivity and dynamic specificity. This is because as demonstrated by Xu and O'Quigley [8], it is easier to estimate the distribution of biomarker conditional on time-to-event by summing up all the survival rates of diseased subjects. However, as shown by Rizopoulos and Kolamunnage-Dona and Kamarudin [9,10], incorporating joint modelling between longitudinal measurements of the biomarker and the time-to-event can further improve the accuracy of the survival model. Joint modelling combines the linear mixed effect of the longitudinal measurements with the survival risk from Cox regression model and has been shown to be methodologically sound and accurate in describing the survival rate of each individual subject. However, due to the model's complexity, substantial time allocation is required for fitting process and when extending this model to perform time-dependent ROC analysis. Cost and computational power might be the main factors restricting the applicability of joint modelling in time-dependent ROC production.

This paper provides a parametric alternative for researchers to produce the time-dependent ROC curve using baseline biomarker value. The model that we proposed can yield a smooth time-dependent ROC curve without the needs to use kernel smoothing technique. In addition, we can adopt this model to describe any dataset regardless of sample size once all parameter estimates are obtained. Using a similar joint modelling theoretical framework, we can reduce the model to accommodate single biomarker value from each subject and estimate the ROC curve under the cumulative sensitivity and dynamic specificity definition. In Section 2, we provided an overview for researchers to conduct time-dependent ROC analysis given that we know the true distribution for each variable of interest. Parametric proportional hazard (PH) model and Bayes theorem are adopted for creating the joint distribution. The Bayesian approach is employed for parameter estimation and adaptive multidimensional integration is used to produce time-dependent ROC curve. In Section 3, some simulations to study the impact on the AUC trend when the association parameter between the biomarker and time-to-event is varied are performed. The AUC is subsequently calculated using the trapezoidal rule. Comparison with nonparametric and semiparametric

techniques is discussed before applying the outlined procedure to generate the time-dependent ROC curve with real-world data in Section 4. Future directions of the work will be discussed in Section 5.

2 Methodology of Research

For this section, we will revisit the definitions of cumulative sensitivity and dynamic specificity used in this study. A brief explanation on the methodology to generate random data using PH model with skewed normal biomarker and skewed normal time-to-event is provided. Subsequently, we will preview the computations needed to conduct parameter estimation and produce the time-dependent ROC curve.

2.1 Cumulative sensitivity, dynamic specificity, and time-to-event ROC

In time-dependent ROC analysis, several definitions for sensitivity and specificity have been outlined throughout the literature. A thorough discussion of the development of methodologies to perform the time-dependent ROC analysis has been made recently [11]. However, as introduced by Pepe et al.[12], we decided to proceed with cumulative sensitivity and dynamic specificity. For studies that rely on baseline (single) value of each subject, the biomarker measurement is commonly observed at the starting or end time point of the study. Assuming the baseline biomarker X follows a skewed normal distribution as introduced by Azzalini et al. [13], we can find the survival density conditional on biomarker for each subject using the parametric proportional hazard model [14–16]. The density function is written as,

$$f(t_i \mid x_i) = h_0(t_i)e^{\beta x_i}[S_0(t_i)]^{e^{\beta x_i}} \tag{1}$$

where h_0 and S_0 are respectively the baseline hazard and survival function of a skewed normal distribution while β is the association parameter that links information between the biomarker and time-to-event. We used the conditional probability to obtain the joint density g which is,

$$g(t_i, x_i) = f(t_i \mid x_i) \times p(x_i) \tag{2}$$

such that p is the density function for the baseline biomarker. Studies that use the baseline measurements commonly conduct the survival analysis after some data collection period. Thus, the parameter estimates that we will obtain are specifically for our observation time, s. We assume these estimates are fixed over time. As discussed in Pepe et al. [12], at a specific observation time point s, a cumulative sensitivity is the probability of biomarker that is greater than a threshold m, given that the time-to-event is less than or equal to the next t time unit. A dynamic specificity is defined as the probability of biomarker being lesser than or equal to a threshold m, given that the time-to-event is greater than the next t time unit. We can compute both quantities respectively as follows,

$$Se^{C}_{t,s}(m) = P(X > m \mid s < T \leq s + t) = \frac{\int_s^{s+t}\int_m^{\infty} g(u,v)\, du\, dv}{\int_s^{s+t}\int_{-\infty}^{\infty} g(u,v)\, du\, dv}, \tag{3}$$

$$Sp^{D}_{t,s}(m) = P(X \leq m \mid T > s + t) = \frac{\int_{s+t}^{\infty}\int_{-\infty}^{m} g(u,v)\, du\, dv}{\int_{s+t}^{\infty}\int_{-\infty}^{\infty} g(u,v)\, du\, dv}. \tag{4}$$

Denote when t approaching zero, the sensitivity is defined as incident sensitivity. We can generate the ROC curve by plotting all the computed sensitivity and one minus specificity pairs.

2.2 Random data production using proportional hazard model

Random data of the biomarker X can be generated using any statistical package in R, while the time-to-event data for each sample is produced using the following relationship,

$$\begin{aligned} S(t_i|x_i) &= S_0(t_i)^{e^{x_i\beta}} \\ U_i &= S_0(t_i)^{e^{x_i\beta}} \\ \frac{ln(U_i)}{e^{x_i\beta}} &= ln\big(S_0(t_i)\big) \\ {S_0}^{-1}\left(e^{\frac{ln(U_i)}{e^{x_i\beta}}}\right) &= t_i. \end{aligned} \tag{5}$$

The second step is performed by employing the integral probability theorem which suggests the cumulative distribution of a continuous function is uniformly distributed [17]. After obtaining all true time-to-event data, we can use the algorithm proposed by Ramos et al. [18] to generate censoring time. However, as the method to generate random censoring time requires the usage of $Uniform(0,\lambda)$ where λ is the maximum limit for time, we will use a modified algorithm to enable us working with continuous distribution that have domain support which lie on the real line number. We select π% of the observed time to have status 0 and $1-\pi$% of the observed time as status 1, where π corresponding to the right censoring rate of our interest. We can use the 'sample' function in R to randomly select the index row of the true time-to-event that will be censored. The following algorithm can be applied in R to generate random data of PH model that uses the skew normal distribution.

Algorithm 1: Random data generation

Generate n random samples from a $Uniform(0,1)$ distribution
Generate n random biomarker measurement from a continuous distribution
while (i < n)
    compute $t_i\ through\ {S_0}^{-1}$
    $i \leftarrow i + 1$
end
Generate π% random index number
if (row = index)
    $\delta_i \leftarrow 0$
else
    $\delta_i \leftarrow 1$
end

2.3 Parameter estimation and AUC computation

In this paper, we assumed the data mix with right censoring data. Adjustment on the likelihood function is required. The likelihood is written as,

$$\begin{aligned} L(\boldsymbol{\theta}; t_1, \dots, t_n, x_1, \dots, x_n) &= \prod_{i=1}^{n} f(t_i, x_i)^{\delta_i} \left(\int_{t_i}^{\infty} f(u_i, x_i) du_i\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(t_i, x_i)^{\delta_i} S(t_i|x_i)^{1-\delta_i} \end{aligned}$$

with $\boldsymbol{\theta}$ representing the model parameter, $\delta_i = 0$ for censored time-to-event and $\delta_i = 1$ if not censored. To ease the implementation of Bayesian inference computation, we used the log-likelihood function, which is written as,

$$l(\boldsymbol{\theta}; t_1, \dots, t_n, x_1, \dots, x_n) = \sum_{i=1}^{n} \delta_i \, log\big(f(t_i, x_i)\big) + (1 - \delta_i) \Big(log\big(S(t_i|x_i)\big)\Big)$$

Recalling Bayes theorem, the posterior distribution $P(\boldsymbol{\theta} \mid x, t)$ of a vector of parameters $\boldsymbol{\theta}$ is defined as,

$$P(\boldsymbol{\theta} \mid x, t) = \frac{P(x, t \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta})}{\int_{\theta} P(x, t \mid \boldsymbol{\theta}) \, P(\boldsymbol{\theta}) \, d\boldsymbol{\theta}}$$

with $P(x, t \mid \boldsymbol{\theta})$ as the log-likelihood function and $P(\boldsymbol{\theta})$ as the prior distribution of vector $\boldsymbol{\theta}$. The denominator of equation above is a constant, while the numerator varies based on $\boldsymbol{\theta}$. Thus, we only need to find the $\boldsymbol{\theta}$ that maximise the numerator part. Assuming the parameters are independent of each other, the equation above can be written as

$$P(\boldsymbol{\theta} \mid x, t) = P(x, t \mid \boldsymbol{\theta}) \times \prod_{i=1}^{k} P(\theta_i).$$

where k is the count for number of parameters that we need to estimate. To perform Bayesian inference, we rely on the Stan program, as several functions related to the skewed normal have been developed and can be directly used. We use the log skewed normal function and the log complementary cumulative distribution function of skewed normal. For all parameters with real line support domain, the $Normal(0,3)$ is used as the prior distribution. Meanwhile, the $Gamma(0.5,0.5)$ prior distribution is used for parameters that lie within positive real line support domain. We modify the algorithm shown in Kelter [19], which uses the Stan model and R program to estimate the parameters for exponential proportional hazard survival model.

To solve equations (3) and (4), numerical method which is the adaptive multivariate integration from the 'cubature' package of R is employed due to its fast convergence and ease of implementation. The package, which is written by Narasimhan [20], acts as a wrapper to run numerical algorithms that used deterministic and Monte Carlo methods. Using the tensor product of Clenshaw-Curtis quadrature rules, the adaptive integration method repeatedly doubles the degree of the quadrature rules until convergence is achieved [21]. After plotting all sensitivity and specificity rates, trapezoidal rule is used to calculate the area under the ROC curve.

3 Simulation studies and results

For simulation studies, we compared the impact of different sample sizes on the time-dependent ROC curve produced by the PH model (parametric), Kaplan-Meier (nonparametric) [2], and smoothROCtime (semiparametric) techniques [6]. Random biomarker and time-to-event data with sample sizes of 35, 60, 90 and 100 are generated with a censoring rate of 0.1. We used the parameters in Table 1 to conduct the sample size study.

Table 1. Parameters used for simulation to study impact of sample size.

Parameters	μ_X	σ_X	α_X	μ_T	σ_T	α_T	β
True Value	0	1	1	0	1	1	0.5

For all sample size, the Bayesian approach achieve convergent estimator as the number of effective sizes are greater than 100 and the 'Rhat' are closed to one. However, as shown in Table 2, it is observed that some estimators exposed to high bias such as the skewness parameter for time-to-event T and biomarker X. Increasing sample size can reduce the bias for skewness parameter significantly.

Table 2. Bias between estimated and true value.

Parameters	True Value	Bias (sample size)			
		(35)	(60)	(90)	(100)
μ_X	0	0.10	-0.09	0.36	0.18
σ_X	1	0.18	0.01	-0.01	-0.26
α_X	1	0.92	0.70	-0.57	0.61
μ_T	0	1.34	0.75	0.09	-0.42
σ_T	1	0.07	0.33	0.19	0.39
α_T	1	-3.60	-0.80	0.15	1.94
β	0.5	0.04	0.22	-0.06	-0.19

The produced time-dependent ROC curve in Figure 1 shows high volatility for nonparametric ROC curve when the sample size is 35 and 60. As we increased the sample size, the produced time-dependent ROC improved for nonparametric method, but the curve still shows a nonmonotone sensitivity and specificity. This is not the case for parametric and semiparametric methods. However, when comparing the AUC produced from these three methods, we can see that both parametric and semiparametric methods always produced AUC measurements that are lower than the nonparametric method. The underestimation shown might be due to the censoring data or kernel bandwidth selection problem for semiparametric method [4,5,22].

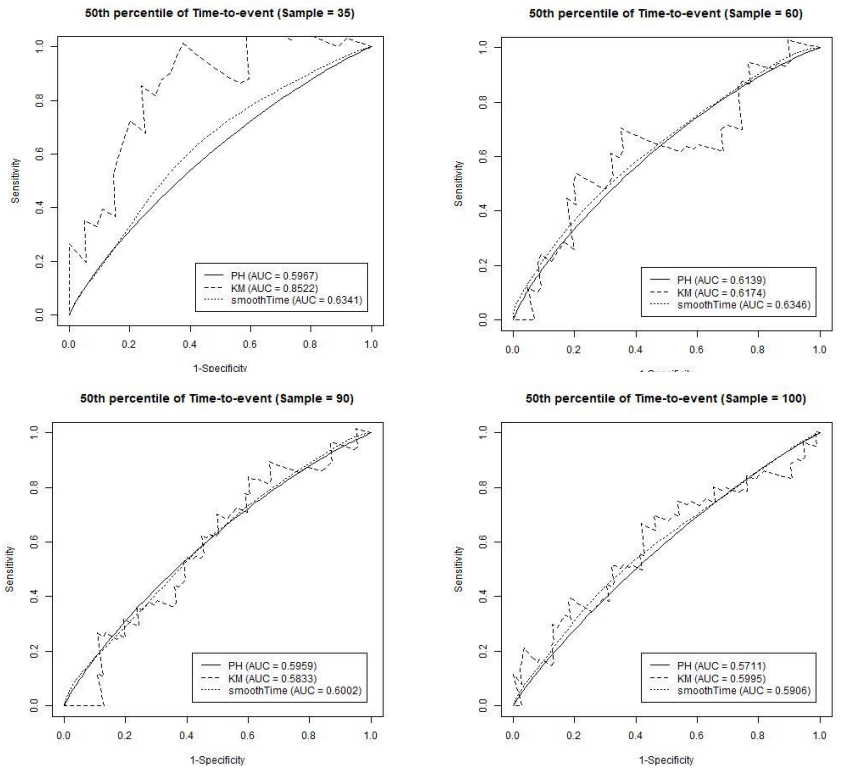

Fig. 1. Time-dependent ROC curve for PH model, Kaplan-Meier (KM) and smoothROCtime (smoothTime).

Next, we studied the impact of changing the association parameter β of the PH model on the AUC trend. The association parameter with values -0.9, -0.5, 0.5 and 0.9 are used to demonstrate a moderate, and strong associations in both negative and positive association between the biomarker and time-to-event data. We produced the time-dependent ROC to assess the accuracy of correctly classifying subjects into cases (diseased) or controls (nondiseased) for the next 0.5-time unit. Thus, for the simulation study, we will compute the following sensitivity and specificity,

$$Se^{C}_{0.5,s}(m) = P(X > m \mid s < T \leq s + 0.5),$$
$$Sp^{D}_{0.5,s}(m) = P(X \leq m \mid T > s + 0.5)$$

with observation time s constituting to nine quantile points of the generated time-to-event T. The AUC of the ROC curve for each observation time s is plotted in a graph. Four combinations of scenario are used to consider all possibilities of skewness for biomarker and time-to-event. After generating the random data for skewed normal biomarker X and skewed normal time-to-event T using the parameter values as shown in Table 3, the skewness for the marginal distribution is determined using the histogram of each individual biomarker and time-to-event. The parameter μ, σ and α are respectively defined as the location, standard deviation, and the skewness of the biomarker X and time-to-event T.

Table 3. Parameters used for simulation to study impact of varying beta on the AUC.

Parameters	μ_X	σ_X	α_X	μ_T	σ_T	α_T
Situation 1	0	1	1	0	1	1
Situation 2	0	1	1	0	1	-1
Situation 3	0	1	-1	0	1	1
Situation 4	0	1	-1	0	1	-1

We bootstrapped the procedure to generate random data. The skewness of marginal distribution of the biomarker and time-to-event is determined through the 'skewness' function from the 'moments' package of R. We observed a right skewed time-to-event irrespective of the situation and association level. However, for the biomarker, the skewness on the marginal is determined by the α. A positive α results in a right skewed biomarker, while a negative α results to a left skewed biomarker. The marginal distribution of each variable is outlined in Table 4 below. We provided examples for the data generated using situations 1 and 3 in Figure 2.

Table 4. Marginal distribution of X and T produced using different parameters combination.

Parameters	μ_X	σ_X	α_X	μ_T	σ_T	α_T	β	X	T
Situation 1	0	1	1	0	1	1	-0.9	Right skew	Right skew
							-0.5	Right skew	Right skew
							0.5	Right skew	Right skew
							0.9	Right skew	Right skew
Situation 2	0	1	1	0	1	-1	-0.9	Right skew	Right skew
							-0.5	Right skew	Right skew
							0.5	Right skew	Right skew
							0.9	Right skew	Right skew
Situation 3	0	1	-1	0	1	1	-0.9	Left skew	Right skew
							-0.5	Left skew	Right skew
							0.5	Left skew	Right skew
							0.9	Left skew	Right skew
Situation 4	0	1	-1	0	1	-1	-0.9	Left skew	Right skew
							-0.5	Left skew	Right skew
							0.5	Left skew	Right skew
							0.9	Left skew	Right skew

Fig. 2. Examples of histogram for biomarker and time-to-event data for Situation 1 and 3.

Figure 3 shows the AUC trend of the time-dependent ROC curve for different association levels β in situations 1 and 3.

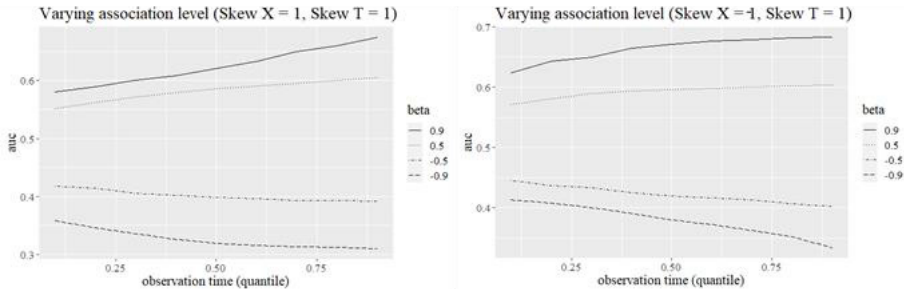

Fig. 3. Time-dependent AUC with a varying association level β for situation 1 and 3 at multiple observation time.

When varying association measurement β, we observed a divergent AUC trend in all situations. For a positive β, the AUC shows an increasing pattern while a negative β shows a decreasing pattern. Increasing AUC trend indicates that at each observation time s, the accuracy of our model to correctly classify for the next 0.5-time unit is improving. Meanwhile, decreasing AUC trend indicates that at each observation time s, accuracy of our model to correctly classify for the next 0.5-time unit is worsening. In situation 3 and 4, we observed a positive β with an increasing AUC trend, but the increasing rate is diminished at a further observation time. One reason that might explain such an observation is because we are using a proportional hazard model that expects to see a low survival rate from subjects with high biomarker measurement but a high survival rate from subjects with a low biomarker measurement. From Table 4, since situations 3 and 4 generate an opposite situation from what should be expected, the performance to accurately separate subject into either class does not greatly improve at a further observation time. From the simulation study, we observed a negative β level that always produces an AUC lower than 0.5, which suggests our model's inefficiency in making a correct classification. The reason for such an occurrence is that a negative association indicates a better ability to survive for a high biomarker measurement. Since our classification model predicts a high biomarker as diseased while a low biomarker as nondiseased, thus we just make a false prediction for both groups. To solve this issue, we need to switch our classification rules. That is to say, the sensitivity should predict the probability of low biomarker as diseased, and specificity should predict the probability of high biomarker as nondiseased. However, this claim warrants further study.

4 Application

The discussed PH model is fitted to the 'wpbc' dataset, which can be found from the 'TH.data' package [23,24]. The dataset contains observations on 198 breast cancer patients treated by Dr. Wolberg in Wisconsin since 1984. The dataset consists of computed measurements of the breast tissue cell nuclei based on a digitized image of a fine needle aspirate [24]. This includes measurement of the perimeter, radius, area, and relevant metrics that describe the characteristics of the cell. We used the measurements of the mean area and perimeter of the cell as the baseline biomarker and the time of breast cancer disease recurrence as the time-to-event. The time-to-event is measured in months. Since measurements of the time and biomarker might be different, we rescaled each variable by dividing the measurement with its corresponding maximum observed value. Rescaled value will transform the support domain of the data between zero to one and can ease computation during parameter estimation and production of the time-dependent ROC curve. The following equations are used,

$$t_i^{new} = t_i / max(t_1, t_2, \dots, t_n)$$
$$x_i^{new} = x_i / max(x_1, x_2, \dots, x_n).$$

Table 5 summarizes the estimated parameters with the 95% credible interval. All parameters achieved convergence since all 'Rhat' values are closed to one, and the number of effective sizes are greater than 100.

Table 5. Estimated parameters using Bayesian inference.

Parameters for mean area of tumor cell						
Parameters	Estimate	Std. deviation	$p_{0.025}$	$p_{0.975}$	Num. effective	Rhat
μ_X	0.31	0.08	0.21	0.56	858	1
σ_X	0.26	0.05	0.17	0.35	1922	1
α_X	2.86	1.85	-0.60	7.11	1511	1
μ_T	0.17	0.16	-0.03	0.59	948	1
σ_T	1.99	0.50	1.19	3.11	2365	1
α_T	8.07	2.44	3.30	12.91	1410	1
β	1.81	0.99	-0.02	3.83	2168	1
Parameters for mean perimeter of tumor cell						
μ_X	0.60	0.09	0.48	0.80	921	1
σ_X	0.17	0.03	0.12	0.24	2061	1
α_X	1.22	1.67	-1.63	4.86	1094	1
μ_T	0.06	0.13	-0.07	0.46	599	1
σ_T	1.47	0.55	0.74	2.82	1273	1
α_T	8.63	2.20	4.38	12.96	2114	1
β	0.17	0.94	-1.40	2.48	924	1

Denote from the estimates we obtained, the skewness parameter α are positives for both biomarkers. This suggests that our dataset is similar to Situation 1 from the simulation study. In addition to the positive estimates of the β for both biomarkers, this indicates that we just observed a situation with high biomarker that leads to lesser time needed for disease to recurring while a low biomarker leads to longer time needed before the disease recurrence. Thus, we can use equations (3) and (4) and expecting to observe increasing AUC pattern for both biomarkers. We generated the ROC curve for 9 observation months and calculated the model's accuracy to correctly classify high (diseased) and low (nondiseased) biomarker for the next one month. As shown in Figure 4, the produced time-dependent ROC curve shows an increasing AUC trend for both biomarkers, indicating an improvement in the model's accuracy performance. Comparable to Figure 3 of the simulation studies, the model's

accuracy when using the mean area of the tumor cell is greater than when using the mean perimeter of the tumor cell because the association β of the former biomarker is higher than the later. It is noteworthy to mention that we might observe small AUC improvement when using the perimeter of the tumor cell as our classifier because the perimeter has relatively small or no association with the time-to-event. Figure 4 shows that using the mean area of the tumor cell is sufficient to segregate subjects into case or control group for the next one month when compared with the mean perimeter.

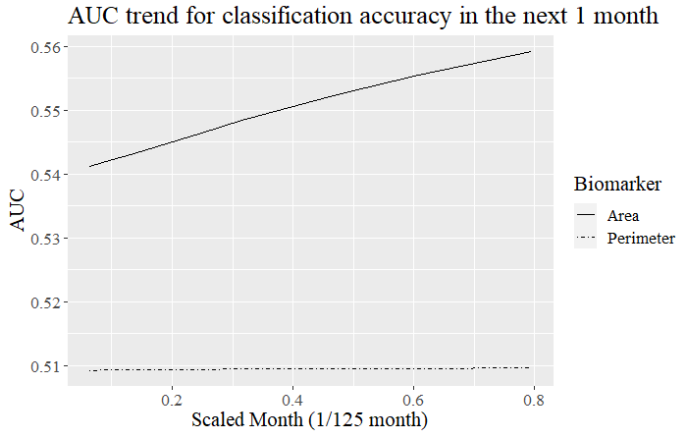

Fig. 4. AUC estimates for mean area and perimeter of tumor cell.

Although the association parameter for the mean area of tumor cell is greater than what we have simulated previously, the AUC for the mean area of tumor cell is still not consider as high enough as it supposed to be when compared to Figure 3. One reason might be due to the distinct skewness differences between the biomarker and time-to-event. In the simulation studies, the ratio of skewness between the biomarker is 1:1 whereas the ratio for our real dataset is estimated about 1:3. In accordance with this, investigation to see how skewness differences can impact the AUC will be an interesting topic to discover.

5 Discussion

In this paper, we have demonstrated a parametric technique that can be applied to measure the model accuracy performance over time. The parametric time-dependent ROC method, which uses a proportional hazard model and skewed normal distribution to connect the biomarker and time-to-event information might assist researchers in conducting comparison tests and identifying distinct variables that can provide high classification accuracy over time. This study is important to initiate the study on factors impacting the AUC trend on the time-dependent ROC curve. For instance, the parameter that controls association between the biomarker and time-to-event can be simulated to see the influence it brings to the performance of the model's predictive ability. Subsequently, this might aid researchers to obtain initial insights on each marker of interest that will be compared to each other.

Although using parametric method requires strong assumptions on the dataset, this technique might produce superior results when a small sample size is collected in a study. The simulation on the sample size shows that the nonparametric method requires a large sample size before being able to produce a good ROC curve. A small sample size usually led to an unusual, jagged ROC curve produced through the nonparametric method. Meanwhile for the semiparametric method, extra consideration is needed when using semiparametric methods that use kernel function to estimate the survival rate.

For future work, we would suggest extending this parametric technique to accommodate a time-varying association parameter. As discussed in Zhang et al. [25], it is noticeable that for each observation time point, we will observe different risk set which results to a change in the association value accordingly. Allowing the association parameter β to be flexible and time-varying enables us to determine which observation time point will have a high and low association value. Although joint modelling is computationally expensive, as shown in the works of Rizopoulos and Kolamunnage-Dona and Kamarudin [9,10], considering longitudinal measurement on the covariates might also elevate the accuracy of the study, thus strengthening the research findings. Using joint model to link the longitudinal measurement of individuals with the population also might prevent underestimation issues which are important in survival analysis [9].

When considering multivariable studies using parametric approach, problem arise when determining numerical procedures that will be used to conduct the ROC computation. Monte Carlo and adaptive integration usually require substantial computational power and effort, thus additional studies that can bring fast and stable computation are feasible. We also have observed an outstanding focused on using the copula function as computational alternatives towards the joint distribution. As shown in the works of Escarela et al., Zhang et al., and Ganjali and Baghfalaki [26–28], the copula function is able to model the joint distribution while considering all censorship and truncation data, thus providing a convenient alternative as no information appears to be lost.

In this paper, we also have seen how the time-dependent ROC can be applied to compare predictive performance for different biomarkers. Since continuous monitoring of model performance is essential in many applications, time-dependent ROC facilitates the ongoing assessment of a model's predictive power, allowing for timely adjustments or interventions if the performance degrades or improves. We also provided an algorithm to generate random samples from the PH model that is easy to be done and might offer practical utility.

Software availability

The R codes and functions to run simulation and real data application are accessible upon request.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Universiti Teknologi Malaysia (UTM), UTM Fundamental Research (UTMFR) Vote No: Q.J130000.3854.22H03.

References

1. R. Etzioni, M. Pepe, G. Longton, C. Hu, G. Goodman, Med. Decis. Making, **19**(3), 242–51 (1999)
2. P.J. Heagerty, T. Lumley, M.S. Pepe, Biometrics, **56**(2), 337–44 (2000)
3. I. Unal, Comput. Math. Methods Med., **1**, 3762651 (2017)
4. P. Martínez-Camblor, J.C. Pardo-Fernández, Stat. Methods Med. Res., **27**(3), 651–74 (2018)
5. K.M. Beyene, A. El Ghouch, Stat. Med., **39**(24), 3373–96 (2020)

6. S. Díaz-Coto, P. Martínez-Camblor, S. Pérez-Fernández, Comput. Stat., **35**(3), 1231–51 (2020)
7. Y. Zheng, P.J. Heagerty, Biostat., **5**(4), 615–32 (2004)
8. R. Xu, J. O'Quigley, J. Nonparametric Stat., **12**(1), 83–107 (1999)
9. D. Rizopoulos. Joint models for longitudinal and time-to-event data: With applications in R, CRC Press (2012)
10. R. Kolamunnage-Dona, A.N. Kamarudin, Res. Methods Med. Health Sci., **2**(2), 51–60 (2021)
11. A.N. Kamarudin, T. Cox, R. Kolamunnage-Dona, BMC Med. Res. Methodol., **17**(1), 53 (2017)
12. M.S. Pepe, Y. Zheng, Y. Jin, Y. Huang, C.R. Parikh, W.C. Levy, Lifetime Data Anal., **14**(1), 86–113 (2008)
13. A. Azzalini, Scandinavian J. Stat., 171 – 178 (1985)
14. J. Kalbfleisch, R. Prentice, The Statistical Analysis of Failure Time Data, John Wiley & Sons (1980)
15. G. Brostrom, Event History Analysis with R, Chapman & Hall/CRC, 85–112 (2012)
16. S.K. Khosa, Parametric Proportional Hazard Models with Applications in Survival Analysis, PhD diss., University of Saskatchewan, (2019)
17. J.E. Angus, SIAM Review, **36**(4), 652–4 (1994)
18. P.L. Ramos, D.C.F. Guzman, A.L. Mota, F.A. Rodrigues, F. Louzada, Sampling with censored data: a practical guide, arXiv preprint arXiv:2011.08417 (2020)
19. R. Kelter, BMC Med. Res. Methodol., **22**(1), 58 (2022)
20. B. Narasimhan, M. Koller, S.G. Johnson, T. Hahn, A. Bouvier, K. Kiêu, et al., cubature: Adaptive Multivariate Integration over Hypercubes, available from: https://bnaras.github.io/cubature/ [cited 2023 Dec 4]
21. C.W. Clenshaw, A.R. Curtis, Numer. Math. (Heidelb.), **2**(1), 197–205 (1960)
22. S. Díaz-Coto, N.O. Corral-Blanco, P. Martínez-Camblor, IJB, **17**(1):117–37 (2021)
23. W.N. Street, O.L. Mangasarian, W.H. Wolberg, *An Inductive Learning Approach to Prognostic Prediction*, in Machine Learning Proceedings 1995, 522–30 (1995)
24. P. Bühlmann, T. Hothorn, Statistical Science, **22**(4) (2007)
25. Z. Zhang, J. Reinikainen, K.A. Adeleke, M.E. Pieterse, C.G.M Groothuis-Oudshoorn, Ann. Transl. Med. **6**(7), 121–121 (2018)
26. Z. Zhang, C. Charalambous, P. Foster, CSDA, **181**, 107685. (2023)
27. G. Escarela, A.R. Vásquez, G. González-Farías, J.U. Márquez-Urbina, Stat. Methods Med. Res., **32**(6), 1203-1216. (2023)
28. M. Ganjali, T. Baghfalaki, J. Biopharm. Stat., **25**(5), 1077–99 (2015)