Research Article

BIOMEDICAL
Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2024.55.008657

Clustering Algorithm for the Connected Model of Repeated Measurements and Survival Data: Application to HIV Study

Hessah Alzahrani*

Mathematics Department, College of Applied Sciences, Umm AlQura University, Saudi Arabia

***Corresponding author:** Hissah Alzahrani, Mathematical Sciences Department, College of Applied Sciences, Umm AlQura University, Meccah, 24382, Saudi Arabia

ARTICLE INFO

Received: February 02, 2024

Published: February 19, 2024

Citation: Hessah Alzahrani. Clustering Algorithm for the Connected Model of Repeated Measurements and Survival Data: Application to HIV Study. Biomed J Sci & Tech Res 55(1)-2024. BJSTR. MS.ID.008657.

ABSTRACT

Cluster analysis is dividing the individuals into clusters or groups. This job is valuable and helpful to provide facts and information about individuals. Here in this paper, we proposed the cluster analysis for a special model. It is the connected model for repeated measures called longitudinal and survival for individuals or patients. The statistical analysis of the connected model from the longitudinal and survival datasets has become popular recently because it comes together in many medical applications. Here in this study, we utilized two statistical methods, cluster analysis and, connect the two models from longitudinal and survival data. It is beneficial since it gathers information from repeated measurements, and survival responses. The shared random effect term is used joint multivariate Gaussian distribution (longitudinal) and Cox proportional hazard model (survival)for the same patients. Then, the pseudo-liklihood algorithm (clustering methodology) is performed for the joint model to distinguish the clusters or groups of patents. The application is HIV patient's dataset with CD4 + counts responses and time to death (survival data) with some independent variables as gender and drug treatments. We conducted the clustering for S=2 and estimated the parameters from the longitudinal and survival models with and without clustering and compared the estimations. Our results showed the generated clusters are different from each other, the estimation parameters are located around the original estimations (without clustering). It is a helpful methodology to identify distinct groups or clusters from population. Finally, there is a big need for this type of application in medical fields.

Introduction

Many clinical trials applications generate repeated measurements and time to event (survival data). In longitudinal studies the patients are followed over many occasions (repeated measurements) and their data indicates biomarkers. Sometimes these longitudinal data include time to event, for example time to death, Alzahrani [1]. There are many statistical methodologies designed to joint or connect the analysis of the repeated measurements and survival data for some reasons. Here in our study, we conducted the cluster analysis for a group of patients from their longitudinal and survival models. Moreover, cluster analysis is a statistical methodology that seeks to separate subjects into new groups based on increase homogeneity inside each group and heterogeneity between groups. The clustering analysis could be performed for variables or whole models, which include dependent and independent variables, Ilmarinen, et al. [2,3]. Clustering or classification the patients based on joint analysis of longitudinal and survival models could be beneficial to gather more facts and information from the new groups. The joint statistical analysis for longitudinal and survival data together has a wide range of resent applications Ghisletta [4,5]. The joint latent class model can be viewed as clustering the longitudinal and survival data, dividing the population into finite of latent homogeneous subgroups. The latent term model is based on assuming the population are homogeneous latent groups of subjects, Henry [6-8] applied the latent term method for subjects sharing same responses and same risk of event using MLE method via EM algorithm.

MLE through EM algorithm starts to be complicated for models with random effects of higher dimensions. Also, the clustering of the repeated measurements and survival data can be connected by defining the marginal density of the responses also as mixture distribution. Kom rek [9] applied the Bayesian estimation of the finite mixture models to cluster longitudinal and survival outcomes. Clustering is a common method and there are many R packages applicable for these problems. Bruckers, et al. [10] propose clustering using pseudo-likelihood algorithm for multivariate re- peated measurements outcomes and performed the clustering using k-means criteria using the pairwise approach. Their algorithm allocates N observations in clusters or groups based on maximizing their joint models. The cluster criteria are the individual's likelihood contribution. We borrow this idea but for the individual's joint likelihood model as a cluster criterion. We accommodate his algo- rithm, but for connected model from repeated measurements(longitudinal) and survival datasets. The goal is an attempt determines clusters or groups of patients based their characteristics from repeated measurements and survival data. The clustering algorithm will be based on connected models of the change in the longitudinal responses of a subject and the risk of the survival event. The repeated measure's part and time to event part are conditionally independent given the subject- specific intercept and slope (latent variables).

The main interesting point in this study is connecting the repeated measurements and survival datasets, since they were obtained from the same patients. This natural correlation may lead to new conclusions from the new unknown groups. Modeling the longitudinal and survival data is familiar in real life, Sweeting, and Thompson [11,12]. We applied clustering algorithm in a suitable application, which the is HIV study. AIDS clinical trial is an appropriate example in which the information of the patients is obtained over many occasions. Here in the HIV study, we compared two treatments, didanosine(ddl) and zalcitabin(ddc). The response is the longitudinal outcome, which is the number of CD_4 cells per cubic millimeter of blood, obtained over many occasions to measure the progression of the AIDS disease. However, the time to death (survival data) has a logical relationship to the CD_4 biomarker in the longitudinal model. Classification the AIDS patients based on their longitudinal and survival data is an interesting research idea to evaluate the two treatments, ddl and ddc. This paper has the following structure, the clustering algorithm in section two is reviewed. Then, section three contains the application of the HIV study where the clustering algorithm is applied. Section 3 contains the study description, the proposed model, and the results. Finally, the conclusion is in section 4.

The Clustering Method

Let $i = 1, 2, \ldots N$ is the number of observations and $j = 1, 2, \ldots j_i$ is the number of occasions. For easier notations, we will refer to j_i to j , assuming all patients have the same number of occasions. $time_{ij}$ is a time of subject i at j occasion. Then, the outcomes can be seen as multivariate Gaussian distribution for the longitudinal responses:

$$Y_{ij} \mid a_i, b_i = \mathrm{X}_{\mathrm{it}}\beta + a_i + b_i time_{ij} + \varepsilon_{ij} \quad (1)$$

The Cox proportional hazard ration is:

$$\mathrm{h}_{\mathrm{i}}(\mathrm{t}) = \mathrm{h}_{\mathrm{o}}(\mathrm{t}) \exp(\mathrm{W}_I^T \Upsilon + \mathrm{a}_{\mathrm{i}}\delta_1 + \mathrm{b}_{\mathrm{i}}\delta_2) \quad (2)$$

where $\varepsilon_{ijk} \sim \mathrm{N}(0, \sigma)$ and $\substack{ai \\ bi} \sim \mathrm{N}\left(\substack{0 \\ 0}, \substack{\sigma^2 \;\; \sigma_{ab} \\ \sigma_{ab}^a \;\; \sigma_b^2}\right)$. δ_1 and δ_2 are scalars. For survival part, $\mathrm{h}_{\mathrm{i}}(\mathrm{t})$ is hazard of death of occasion t conditional on $\substack{a_i \\ b_i}$, D_i is time of death, C_i is censoring time, $T_i = min(C_i, D_i)$ is observed time, w_i is the covariates vector for individual in the hazard model, and $\Delta_i = \mathrm{I}(\mathrm{D}_{\mathrm{i}} < \mathrm{C}_{\mathrm{I}})$. The basic idea of the clustering algorithm is using the maximum likelihood based on joint models for simultaneously analyzing longitudinal and survival data. The latent variables $\substack{a_i \\ b_i}$ are used to link longitudinal and survival submodels, Morrell [13]. b_i is reflecting the rate of change of subject specific mean over time. The cluster criteria is the individual's joint likelihood from the longitudinal and survival models:

$$\log^{\int} f(\mathrm{Y}_{\mathrm{it}}|\mathrm{a}_{\mathrm{i}}, \mathrm{b}_{\mathrm{i}}) f(\mathrm{T}_{\mathrm{i}}|\mathrm{a}_{\mathrm{i}}, \mathrm{b}_{\mathrm{i}}) f(\mathrm{a}_{\mathrm{i}}, \mathrm{b}_{\mathrm{i}}) \mathrm{d}[\mathrm{a}_{\mathrm{i}}, \mathrm{b}_{\mathrm{i}}]$$

Assuming ϑ vector is containing all the parameters from the repeated measurements model, the survival model, and the variance covariance of the random effects. Then, we will use maximum likelihood estimation to estimate ϑ , Song, et al. [14,15]. The clustering approach is based on the likelihood framework. It performed on the following steps:

1. Assume the number clusters S=2 and randomly divide the observations into S clusters.
2. 2- Run the joint modeling of the longitudinal and survival for each cluster separately.
3. Iterate the following steps (a to c) until no observation's switches cluster anymore.

a. Change the cluster assigning for each observation to the other clusters and compute their likelihood depending on the parameters for each cluster.

b. For each observation, compare the likelihood for each cluster and reclassify it for the cluster that has maximum likelihood.

c. Apply the joint modeling of longitudinal and survival for each cluster.

Application to HIV study

Study Description and Models

Starting by introducing the HIV disease, it is a virus attacks the immune system. It results in destroying the CD_4 cells, which are the white blood cells in our immune system. It gradually declines the count of CD_4 and breaks down a patient's immune system. When a patient lives with HIV without any treatment, he will be vulnerable to infections and diseases. Thus, HIV disease progression is delayed when there is a high amount of CD_4 cells. The count of CD_4 cells is a primary indicator of HIV disease. This study belongs to the Community Programs for Clinical Research on AIDS (CPCRA). There were 467 patients who were diagnosed with HIV infection. It was performed in accordance with relevant guidelines and regulations and consents were obtained from all patients. Also, Informed consent has been obtained for all participants in the study. The National Institute of Allergy and Infectious Diseases (NIH) sponsored the CPCRA. The HIV study was performed in accordance with relevant guidelines and regulations to NIH institution (Abrams, et al. [16]). These HIV patients are assigned randomly to get the study treatments are didanosine(ddI) or zalcitabine(ddC), starting by 230 patients in ddI group and 236 in ddC group. The non-missing patients over the five time points in ddI is (230,182,153,102,22) while it is (236,186,157,123,14) in ddC group. It happens in the longitudinal studies to have an increase in the missingness rate over time (dropout) due to many causes such as lack of communications or cure of disease De Gruttola [17-19].

The main outcome is the CD_4 count, which is recorded at the study entry is measured at 6, 12 & 18 months. However, the time to death or censoring is measured for each patient. The dataset is a combination of repeated measurements(longitudinal) and survival data. In this study, Yij denotes the square root of CD_4 count and the independent variables are included in Table 1:

Table 1: Covariates Variables.

Covariate	Definition
Drug	(1) The subject received didanosine (ddI),(0) The subject received zalcitabine (ddC).
Gender	(1) Male, (-1) Female.
Prev	(1) The subject reported previously having infection of AIDS ,(-1) no infection of AIDS disease previously.
Stratum	(-1)The subject has AZT intolerance, (-1) no AZT intolerance .

The linear random effects model for square root CD_4 count is specified as

$$Y_{ij}|a_i, b_i = \beta_{11}+\beta_{12}time_{ij}+\beta_{13}t_{ij}Drug_i+\beta_{14}Gender_i+\beta_{15}Prev_i+\beta_{16}Stratum_i+a_i time_{ij}+b_i \quad (3)$$

The Cox proportional hazard ration is:

$$h_i(t)=h_o(t)\exp\left(\beta_{21}+\beta_{22}Drug_i+\beta_{23}Gender_i+\beta_{24}Prev_i+\beta_{25}Stratum_i+a_i\delta_i+b_i\delta_2\right) \quad (4)$$

Our main goal is to cluster to HIV patients into two groups (S=2) based on the association among CD_4 count, survival time, drug group, gender, AIDS diagnosis at baseline (an indicator of disease progression status), and stratum, accounting for all relevant correlations and subject- specific random effects. Since the survival time for each patient in the study is assumed to follow its own hazard function $h_i(t)$, we assume the survival time for the i^{th} subject follows exponential distribution exponential distribution, $t\sim\exp(\mu_i(t))$, where

$$\log(\mu_i(t))=\beta_{21}+\beta_{22}Drug_i+\beta_{23}Gender_i+\beta_{24}Prev_i+\beta_{25}Stratum_i+a_i\delta_i+b_i\delta_2 \quad (5)$$

The Results

To get a better comparison view, we start by conducting the connected modeling analysis for re- peated measurements and survival datasets without clustering. Then, we performed the clustering methodology in that is described in section2 for the same joint model using SAS software. Assuming the number of clusters is two (S=2), then our methodology divided the HIV patients into two groups from their connected model of the longitudinal and survival data. This clustering is carried out by using a 3-steps process to ensure the best classification from their connected model. Now, we will compare the results of parameter estimations with and without clustering. All these results are from the joint model from the repeated measurements(longitudinal) and survival data of HIV patients. Starting from Table 2, we have the description statistics for the covariates and outcomes. The main outcome for the longitudinal model is the mean of the square root of CD_4 counts and time to death (the survival data). From these two outcomes, we see group 2 has better results, longer time to death (13.06) and higher count of CD_4 over occasions=0, 2 and 6 months are (mean=13.06; SD=5.17), (mean=7.73; SD=4.81), (mean=8.33; SD=5.11), unless at occasion 12 and 18 months where the missingness rate is increased. For the results of covariates, the clustering did not make big difference over the two groups. The key point here is the readings of the longitudinal and survival outcomes for all patients are located between the estimations of the two groups.

Table 2: Descriptive statistics of demographic and clinical variables.

	All patients	Group 1	Group 2
Variable	# patients(%)		
N	486	180	287
Death(1)	188(40.26%)	106(22.70%)	82(17.56%)
Gender(male)	422(90.36%)	163(34.90%)	259(55.46%)
Drug(trt)	230(49.25%)	96(20.56%)	134(28.69%)
Prevoi(1)	302(65.74%)	159(34.05%)	148(31.69%)
Stratum(AZT)	175(37.45%)	79(16.92%)	96(20.56%)
variable	mean(SD)		
t2death(t)	12.63(4.94)	11.95(4.48)	13.06(5.17)
CD4 at time=0	7.13(4.71)	6.17(4.39)	7.73(4.81)
CD4 at time=2	7.34(5.23)	6.08(4.89)	8.33(5.11)
CD4 at time=6	6.59(4.94)	5.70(4.99)	7.12(4.84)
CD4 at time=12	7.03(5.27)	7.70(5.73)	6.88(5.16)
CD4 at time=18	6.54(4.68)	7.32(4.96)	6.322(4.67)

In Table 3 the parameter estimations of the longitudinal model for all patients together and after conducting clustering into two groups. The estimated average of regression coefficient of time covariate for all patients is -0.1668 while its estimation for group 1 is -0.1868 and -0.1618 for group 2. The regression coefficient of Prev covariate, diagnosis of AIDs before stratum, is also significant at -2.3152 before clustering procedure, -2.1774 for group 1, and -2.2064 for group 2.

After clustering the patients, we figured out some points. It seems the regression coefficients after clustering are around (less and more) the estimations before clustering. Also, the regression coefficients of Gender covariate are statistically significant in group 1 and group 2 while it was not significant before clustering the patients into two groups. In this study there are 90.36% are male which makes sense to have significant parameter estimation. Table 4 has the survival model regression coefficients for groups 1 and 2. The Gender coefficient estimation has significant estimation on group 2 (95%CI:, 0.4873; p-value=0.0009) where it was not significant before clustering the patients. However, Figure 1 presents Kaplan-Meier survival plots of group1 versus group 2. Looking for the first 7 months, approximately one month after the baseline, group2 survival outcome has significantly better results than group 1 survival outcome. In Figure 2, the survival curves show the differences between the two types of treatments didanosine(d-dI) and zalcitabine (ddC)for each group separately. The Kaplan-Meier survival curves of group2 generally still has higher results for both treatments than group1. However, the survival outcomes of the two types of drugs in group1 has similar curves, but in group2 the treatment didanosine(ddC) has higher survival outcome than zalcitabine (ddI). We conclude the clustering procedure divides the patients into two really distinct groups.

Table 3: Estimations of coefficients, SD and p-value from the longitudinal models for all patients, group1 and group2.

	All patients			Group 1			Group 2		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	8.0129	0.3511	<.0001	9.3716	0.5114	<.0001	7.1211	0.4633	<.0001
time	-0.1668	0.02038	<.0001	-0.1868	0.0425	<.0001	-0.1618	0.02332	<.0001
time*Drug	0.02998	0.02891	0.3003	0.08011	0.05532	0.1497	0.009227	0.03426	0.7879
Gender	-0.1582	0.3249	0.6265	-1.8742	0.6147	0.0027	1.0011	0.4313	0.0208
Prev	-2.3152	0.2382	<.0001	-2.1774	0.5692	0.0002	-2.2064	0.3434	<.0001
Stratum	-0.1309	0.2352	0.578	-0.5397	0.2905	0.0651	0.2317	0.3593	0.5194

Table 4: Estimation of coefficients, SE and p-value from Cox survival models for all patients, group1 and group2.

	All patients			Group 1			Group 2		
Parameter	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	3.702	0.1619	<.0001	4.3248	0.5372	<.0001	3.6375	0.1974	<.0001
Drug	-0.208	0.1464	0.1553	-0.1539	0.1952	0.4304	-0.3369	0.2227	0.1303
Gender	0.1694	0.1226	0.167	0.0391	0.3114	0.9002	0.4873	0.1471	0.0009
Prev	-0.6195	0.1132	<.0001	-1.4253	0.5285	0.007	-0.2748	0.1605	0.0868
Stratum	-0.0824	0.0815	0.3115	-0.1791	0.0997	0.0724	-0.1306	0.1495	0.3823

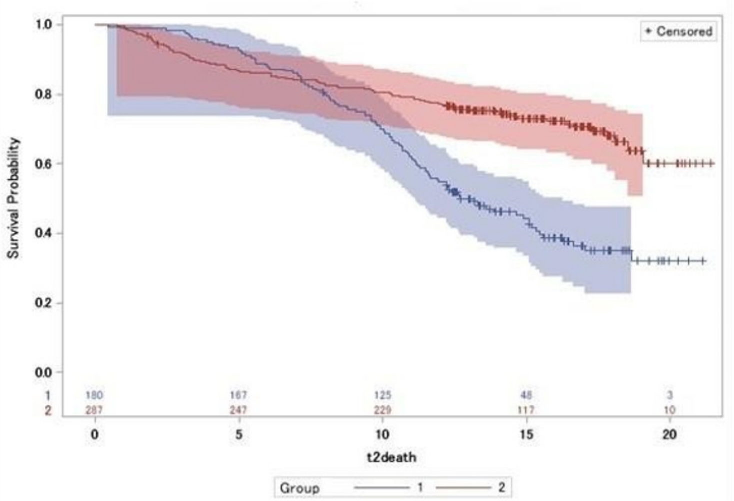

Figure 1: Survival curve by clustering groups 1 and 2.

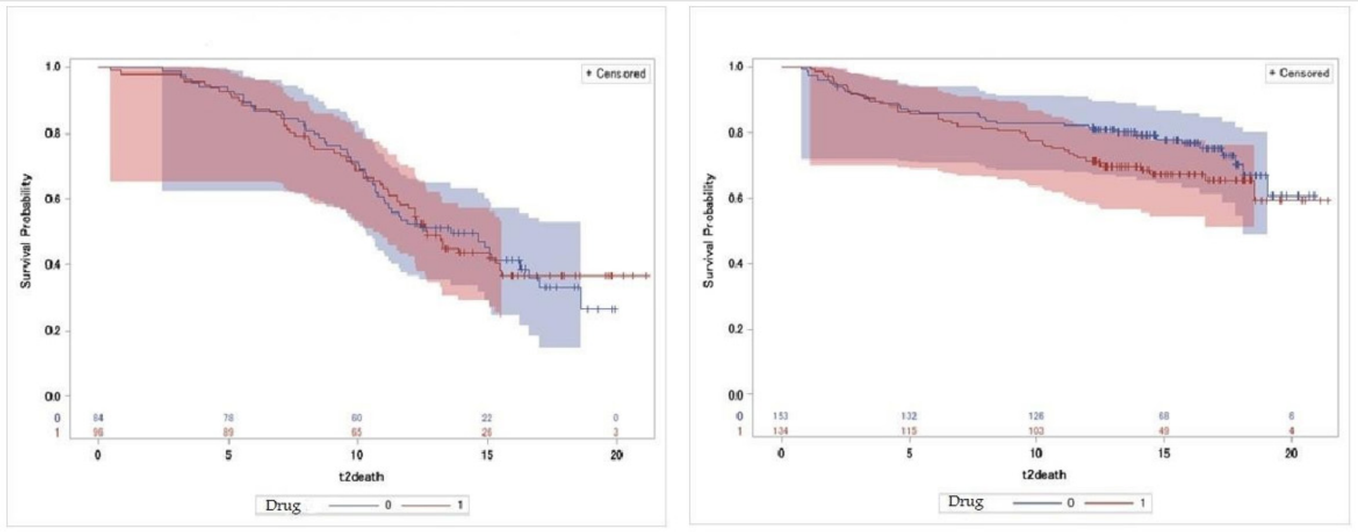

Figure 2: Survival curves by Drug for group 1 and 2, respectively from the left.

Conclusion

In this paper, we build a clustering methodology from the connected models of repeated measurements and survival datasets. The methodology is using the MLE in the clustering algorithm to divide the patients into new groups. The cluster criteria are the joint likelihood from longitudinal and survival models. After some iterative steps, the results are a new classification for S clusters, here we apply it for S=2. The contribution here is identifying new groups of patients based on their repeated measures and survival outcomes. In future, this methodology can be generated in S groups. We found estimation parameters of the new clusters or groups located around the estimation parameters that resulted without doing the clustering procedure, just one group. The application we used is HIV study, consists of patients' reading of CD_4 count, time to death and some covariates. The results distinguish two different groups of patients having different patterns of health associated with longitudinal and survival outcomes. The estimation parameters of the new clusters have deeper facts and information. This classification could help to know the group of that has better outcomes of interest.

References

1. Alzahrani H (2016) Multivariate Binary Longitudinal Data Analysis. PhD thesis, The Florida State University.
2. Ilmarinen P, Tuomisto LE, Niemel O, Tommola M, Haanp J, et al. (2017) Cluster analysis on longitudinal data of patients with adult-onset asthma. The Journal of Allergy and Clinical Immunology: In Practice 5(4): 967-978.
3. Mouli SC, Naik A, Ribeiro B, Neville J (2017) Identifying user survival types via clustering of censored social network data. arXiv preprint arXiv: 1703.03401.
4. Ghisletta P (2008) Application of a joint multivariate longitudinal survival analysis to examine the terminal decline hypothesis in the swiss interdisciplinary longitudinal study on the oldest old. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 63(3): 185-192.
5. DuBois Bowman F, Manatunga AK (2005) A joint model for longitudinal data profiles and associated event risks with application to a depression study. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54(2): 301-316.
6. Henry NW (1983) Latent structure analysis. Encyclopedia of statistical sciences.
7. Clogg CC (1995) Latent class models. In Handbook of statistical modeling for the social and behavioral sciences, pp. 311-359. Springer.
8. Lin H, Turnbull BW, McCulloch CE, Slate EH (2002) Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate- specific antigen readings and prostate cancer. Journal of the American Statistical Association 97(457): 53-65.
9. Kom rek A (2009) A new r package for bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. Computational Statistics & Data Analysis 53(12): 3932-3947.
10. Bruckers L, Molenberghs G, Drinkenburg P, Geys H (2016) A clustering algorithm for multivariate longitudinal data. Journal of biopharmaceutical statistics 26(4): 725-741.
11. Sweeting MJ, Thompson SG (2011) Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. Biometrical Journal 53(5): 750-763.
12. Wulfsohn MS, Tsiatis AA (1997) A joint model for survival and longitudinal data measured with error. Biometrics 53(1): 330 339.
13. Morrell C, Brant L (2000) A two-stage linear mixed-e ects/cox model for longitudinal data with measurement error and survival. In PROCEEDINGS-AMERICAN STATISTICAL ASSOCIATION BIOMETRICS SECTION, UNKNOWN, pp. 198-203.
14. Song X, Davidian M, Tsiatis AA (2002) A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. Biometrics 58(4): 742-753.

15. Albert PS, Shih JH (2010) On estimating the relationship between longitudinal measure- ments and time-to-event data using a simple two-stage procedure. Biometrics 66(3): 983-987.

16. Abrams DI, Goldman AI, Launer C, Korvick JA, Neaton JD, et al. (1994) A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. New England Journal of Medicine 330(10): 657-662.

17. De Gruttola V, Tu XM (1994) Modelling progression of cd4-lymphocyte count and its relationship to survival time. Biometrics 50(4): 1003-1014.

18. Tsiatis AA, Degruttola V, Wulfsohn MS (1995) Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. Journal of the American Statistical Association 90(429): 27-37.

19. Alzahrani H (2018) Missing data analysis for binary multivariate longitudinal data through a simulation study. Biometrics and Biostatistics International Journal 7(2): 103-113.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2024.55.008657

Hissah Alzahrani. Biomed J Sci & Tech Res