Data Analysis of Diabetes Mellitus with Joint Modeling Method

A'yunin Sofro[1*], Muizzatul Mukaromah[2], and Khusnia Nurul Khikmah[2]
[1]Actuarial Sciences Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
[2]Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia
*Corresponding author: ayuninsofro@unesa.ac.id

Received: 1 June 2024
Revised: 24 September 2024
Accepted: 30 September 2024

ABSTRACT – **Diabetes mellitus is a dangerous disease that requires long-term medical treatment. The cause of this disease is high blood sugar levels. If not treated immediately, complications will occur and even cause death. The data is taken from the Indonesia Family Life Survey (IFLS). IFLS is a longitudinal measurement that is performed repeatedly every five years. More data is needed for repeated measures. Therefore, this research needs to be done to accommodate the missing data, and it is assumed that it is missing at random (MAR). This study aims to analyze the causative factors that are thought to affect the recovery time of patients with diabetes mellitus using the joint modeling method. This model is a relationship between event time data and repeated measurement data. The joint modeling method uses a linear mixed model for longitudinal measurements and a Cox proportional hazard model for survival. The variables were taken from IFLS4 and IFLS5 data with 293 observations: measurement time, treatment history, gender, comorbidities, and complications. The results in this study obtained a significant influence, namely the variables of measurement time, gender, and complications, on the recovery time of patients with diabetes mellitus. With the reduced measurement time, the patient has a lower chance of recovering 8.7184 times. The variables of gender also have a lower possibility of recovery of 9.1032 times, respectively.**

Keywords– **diabetes mellitus, Indonesia, joint modeling, longitudinal data, survival data**

I. INTRODUCTION

Diabetes mellitus is a group of metabolic diseases with the leading cause of high blood sugar levels for a long time. If not treated immediately, this disease will cause complications and even death [1]. These complications include possible blindness, kidney failure, stroke, and lower extremity amputation [2]. According to the [3], the number of diabetes cases has continued to increase over the last few decades. There are 9116.03 cases of diabetes mellitus in Indonesia, which causes Indonesia to be the second most common in Southeast Asia [4]. It is estimated that in 2035, the number of diabetes mellitus will increase by 55% due to population growth and urbanization in the world [5]. Based on previous research in [6] mentioned that East Java is one of the provinces with the highest cases of Diabetes Mellitus in Indonesia.

The data is taken from Indonesia Family Life Survey (IFLS) data in 2007, IFLS4 data, and for 2014, IFLS5 data. IFLS is a survey conducted in several areas in Indonesia with several aspects of research. IFLS is longitudinal, with the study's design time every seven years. The iterative approach in IFLS data has been widely developed, one of which is in [7] research investigating longitudinal models on IFLS data. However, this research still has shortcomings. Namely, it has not accommodated the missing data in the observations. The missing data include death, relapse or disease recovery, and response to treatment, for example, not taking medication when the research was conducted. In this study, there is also survival data. To analyze survival data, you cannot use the usual regression method because it will cause bias and cannot handle censored data [8], so the appropriate way to accommodate these deficiencies is joint modeling. This study also presents several explanatory variables that affect diabetes mellitus patients taken from previous studies. Previous research on [9] found that the time of measurement on IFLS4 and on IFLS5 affects. Research on [6] states that the patient's medical history affects. Research on [10], [11], [12] mentions that gender, comorbidities, and complications affect diabetes mellitus.

Joint modeling is a longitudinal and survival approach, whereas the independent covariates joint modeling is a longitudinal and survival approach. Covariate joint models are time-dependent. First, one must distinguish between internal (endogenous) and external (exogenous) covariates. This joint modeling applies to observations when the focus is on survival outcomes to investigate the effect of endogenous time-dependent internal covariates, where measurements in this study were longitudinal. The overall aim of the joint analysis is to study the impact of covariates on longitudinal outcomes, survival outcomes, or both [13]. One example of joint modeling is found in the study of [14], [15], regarding to the recovery time of AIDS (Immunodeficiency Syndrome Study), which correlates the survival time of AIDS patients on the CD4 cell count with longitudinal measurements. Based on the references of several studies above, in this article, a joint modeling approach will be used using a linear mixed model (LME) for longitudinal measurements and a Cox proportional hazard model for survival.

To detect individuals who suffer from diabetes mellitus by examining blood sugar levels. Each measurement is seen from the glycated Hemoglobin (HbA1c) level, which is often used for long-term glycemic control. Leveraging wearable censored and uncensored of longitudinal data, this joint model offers unprecedented insights into the daily lives of diabetes patients, uncovering previously unidentified behavioral patterns linked to disease progression. Existing joint models often focus on estimating blood sugar levels, this novel approach delves deeper, uncovering hidden links between

diabetes mellitus, comorbidities, and complications that potentially paving the way for personalized preventive interventions. So, in this study, further data on diabetes mellitus will be carried out using the joint modeling method using longitudinal and survival data. In this study, we will examine the diabetes mellitus data model in IFLS using joint modeling and analyze what factors affect the recovery time of diabetes mellitus patients in East Java on IFLS data.

II. LITERATURE REVIEW

A. Longitudinal Sub-model

Longitudinal data is data from repeated measurements for a variable with the same number of subjects but at different times [13]. The Linear Mixed Effects (LME) model can be seen as an extension of the standard linear regression model by introducing random effects in the model [16]. The most popular approach in longitudinal data analysis is mixed models in which random effects between subjects are determined to address concerns that there may be unobserved data for each patient or at a certain point [7]. Let n_i be the number of repeated measures for the subject i in a sample of N individuals, and y_i be the n_i dimension vector of repeated measurements. Given linear mixed models as follows [13]:

$$y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + b_{ij} + \varepsilon_{ij} \tag{1}$$

$\boldsymbol{X}_{ij}$ is the $\boldsymbol{n_i} \times \boldsymbol{M}$ vector of covariates for the subject i at time point j. $\boldsymbol{\beta}$ is the $\boldsymbol{M} \times \boldsymbol{1}$ vector of the unknown population parameter called the fixed effect, and b_i is the unknown subject effect, a random effect parameter [17]. In general, the LME model for the longitudinal response y_i can be written as:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b}_i + \boldsymbol{\varepsilon_i} \tag{2}$$

Where $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \dots, y_{in_i})'$ is $n_i \times 1$ repeated measurement vector for the subject, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices, for fixed effect regression coefficient $\boldsymbol{\beta}$, and $\boldsymbol{b}_i$ is random effect regression coefficient. Meanwhile $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{in_i})'$ is a vector of $n_i \times 1$ error measurements in sampling. In this case, we will use a linear mixed effect (LME) model to emphasize the normal distribution of longitudinal results. A similar notation can be written [16]:

$$\begin{aligned} \boldsymbol{y_i}(t) &= m_i(t) + \boldsymbol{\varepsilon_i}(t) \\ \boldsymbol{y_i}(t) &= \boldsymbol{x}_i'(t)\boldsymbol{\beta} + \boldsymbol{z}_i'(t)\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}(t) \\ \boldsymbol{b_i} &\sim N(0, G), \boldsymbol{\varepsilon_i}(t) \sim N(0, \sigma^2) \end{aligned} \tag{3}$$

Where $\boldsymbol{x_i}(t)$ and $\boldsymbol{z_i}(t)$ are design vectors for fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{b_i}$, respectively $\boldsymbol{\varepsilon_i}(t)$ are error terms. Random effects follow a normal distribution with a covariance matrix $\boldsymbol{G}$ (Rizopoulos, 2012). Dealing with measurement error, the observed longitudinal yield $\boldsymbol{y_i}(t)$ is expressed as the sum of the actual longitudinal results on $m_i(t)$ [13]. It is assumed that the sampling error, where $\boldsymbol{\varepsilon_i}$ is independent of $\boldsymbol{b_i}$, is $Cov(\boldsymbol{b_i}, \boldsymbol{\varepsilon_i}) = 0$ with a limit of $0 < \boldsymbol{b_i} < G$ [18]. $\boldsymbol{G}$ is the variance-covariance matrix for random effects across subjects [13]. Matrix $\boldsymbol{G}$ is:

$$G = \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{pmatrix} \tag{4}$$

The interpretation of the fixed effect $\boldsymbol{\beta}$ is assuming p covariate in the design matrix X, the coefficient β_i where $i = 1, \dots, p$ shows the change in the mean $\boldsymbol{y_i}$ when the corresponding covariate $\boldsymbol{X_i}$ increases by one unit, while all other predictors are held constant [13]. Similarly, the random effect, $\boldsymbol{b}_i$, is interpreted as how the subset of regression parameters for the i-th subject deviates from those in the population [7].

B. Survival Sub-model

Survival analysis is a statistical step to analyze data with time until the event occurs. Suppose T_i^* is the survival time, and C_i is the censored time for subject $i = 1, \dots, n$ [13]. T_i denotes the observed failure time for subject $i = 1, \dots, n$, which is defined as $T_i = \min\ (T_i^*, C_i)$ [13]. If the subject is not censored, the patient's survival time has been observed, where $T_i \leq C_i$, and if the subject is censored, then the follow-up has been lost or the subject has died, where $T_i > C_i$ [19]. Suppose δ_i is an indicator of the occurrence of events, then $\delta_i = I(T_i^* \leq C_i)$. The observed survival data are (T_i, δ_i) where $i = 1, \dots, n$ [19]. In general, the Cox proportional hazard model can be written as [13]:

$$h_i(t|\boldsymbol{a_i}) = h_0(t)\exp\{\gamma'\boldsymbol{a_i}\} \tag{5}$$

Where $h_i(t)$ is the individual hazard function i at time t, $h_0(t)$ is the initial unspecified hazard, γ' contains the covariates, and a_i is the vector of the baseline covariates. The primary hazard $h_0(t)$ can be interpreted as a hazard when the covariate values are all zero, for example, $\gamma' = 0$. For instance, $m_i(t)$ shows the longitudinal result of time t for individual i. The actual $m_i(t)$ value is not observed, as will be obtained by measuring error. The following proportional hazard model should be considered to test the relationship between $m_i(t)$ and the hazard in an event.

$$h_i(t|\mathcal{M}_i(t), \boldsymbol{a_i}) = h_0(t)\exp\{\gamma'\boldsymbol{a_i} + \alpha m_i(t)\} \tag{6}$$

Where $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ represents the actual value of the longitudinal covariate for time t. $h_0(t)$ represents a hazard function whose initial value is not defined or individuals with all covariate values of 0. $\boldsymbol{a_i}$ denotes the initial covariate vector, and γ' denotes the vector of regression coefficients corresponding to vector $\boldsymbol{a_i}$. The impact of longitudinal data on survival outcomes at time t is measured with the parameter [13].

C. Basic Joint Model

The Cox model's covariate $\boldsymbol{y_i}(t)$ depends on the measurement error and missing data [16]. For longitudinal data, it is possible to model the path of the observed data y_{ij} based on the simple LME model, then:

$$\boldsymbol{y}_i(t) = X_i(t)'\boldsymbol{\beta} + \boldsymbol{Z_i}(t)'\boldsymbol{b_i} + \varepsilon_i \equiv m_i(t) + \boldsymbol{\varepsilon_i}(t) \quad (7)$$

Following the longitudinal sub-model, $m_i(t)$ is the actual value and other covariates without measuring errors or missing data, and $\boldsymbol{\varepsilon_i}$ is an error measurement.

$$m_i(t) = y_{ij} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} t_{ij} + \boldsymbol{\beta_2 X_{2i}} t_{ij} + b_{0i} + b_{1i} t_{ij} \quad (8)$$

Whereas in the survival sub-model, it is assumed that the event risk is related to the values of the covariates. So, the survival model can be written as:

$$h_i(t) = h_0(t)\exp\left(\gamma' \boldsymbol{a_i} + \alpha[\boldsymbol{\beta_0} + \boldsymbol{\beta_1} t_{ij} + \boldsymbol{\beta_2 X_{2i}} t_{ij} + b_{0i} + b_{1i} t_{ij}\right. \quad (9)$$

The survival and longitudinal models are related in this case, so the inference is needed together. The standard joint model framework for linking survival models and longitudinal models is through shared random effects, often called shared-parameter models. Here, the shared-parameter model can be extended by assuming the longitudinal model and the survival model have a random effects correlation, which can be written as follows [16]:

$$\begin{cases} h_i(t) = h_0(t)\exp\ (\gamma' \boldsymbol{a_i} + \alpha m_i(t)) \\ \boldsymbol{y_i} = \boldsymbol{X_i'\beta} + \boldsymbol{Z_i' b_i} + \boldsymbol{\varepsilon_i} \\ (\boldsymbol{a_i}', \boldsymbol{b_i'})' \sim N(\boldsymbol{0}, \boldsymbol{G}) \end{cases} \quad (10)$$

Where $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$ are random effects correlated with the covariance matrix $\boldsymbol{G}$.

D. Newton Raphson

The application of numerical integration procedures, such as Monte Carlo sampling or to assess $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)}\right)$.

$$\log f\ (\ T_i, \delta_i, \boldsymbol{y_i}, \boldsymbol{b_i}; \boldsymbol{\theta}\ ) = log\ f\ (\ T_i, \delta_i |\ \boldsymbol{b_i}; \boldsymbol{\theta_t}, \boldsymbol{\beta}) + \log f\ (\boldsymbol{y_i}|\ \boldsymbol{b_i}; \boldsymbol{\theta_y}) + log\ f\ (\boldsymbol{b_i}; \boldsymbol{\theta_b}) \quad (11)$$

where,

$$\hat{\sigma}^2 = \frac{\sum_i \int (\boldsymbol{y_i} - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i})'\ (\boldsymbol{y_i} - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i})\ f(\ \boldsymbol{b_i}|\ T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) d\boldsymbol{b_i}}{N} \quad (12)$$

and

$$\hat{G} = \frac{\sum_i \widetilde{v\boldsymbol{b}_\iota} + \widetilde{\boldsymbol{b}_i}\widetilde{\boldsymbol{b}_i}'}{n} \quad (13)$$

where $N = \sum_{i=1}^{n} m_i$ is the total number of observations in the study.

$$\widetilde{\boldsymbol{b}_i} = \boldsymbol{E}\left(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i};\ \boldsymbol{\theta}^{(ic)}\right) = \int\ \boldsymbol{b_i}\, f\left(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i};\ \boldsymbol{\theta}^{(ic)}\right) d\boldsymbol{b_i} \quad (14)$$

$$\widetilde{v\boldsymbol{b}_\iota}\ =\ variance\left(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i};\ \boldsymbol{\theta}^{(ic)}\right) = \int \left(\boldsymbol{b_i} - \widetilde{\boldsymbol{b}_i}\right)^2 f\left(\boldsymbol{b_i}|T_i, \delta_i, \boldsymbol{y_i};\ \boldsymbol{\theta}^{(ic)}\right) d\boldsymbol{b_i} \quad (15)$$

To maximize $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(ic)}\right)$ with respect to $\boldsymbol{\theta}$ can maximize the individual parts with appropriate parameters. The solutions of $\boldsymbol{\beta}$ in the longitudinal submodel and parameters in the survival submodel $\boldsymbol{\theta_t}$ are obtained by Newton Raphson [20]:

$$\widehat{\boldsymbol{\beta}}^{(it+1)} = \widehat{\boldsymbol{\beta}}^{(it)} - \left\{\frac{\delta S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right)}{\delta \boldsymbol{\beta}}\right\}^{-1} S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right) \quad (16)$$

$$\widehat{\boldsymbol{\theta}}^{(it+1)} = \widehat{\boldsymbol{\theta}}^{(it)} - \left\{\frac{\delta S\left(\widehat{\boldsymbol{\theta}}^{(it)}\right)}{\delta \boldsymbol{\theta_t}}\right\}^{-1} S\left(\widehat{\boldsymbol{\theta}}^{(it)}\right) \quad (17)$$

where $\widehat{\boldsymbol{\beta}}^{(it)}$ and $\widehat{\boldsymbol{\theta}}^{(it)}$ are the values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta_t}$ at the current iteration, respectively, $\frac{\delta S\left(\widehat{\boldsymbol{\beta}}^{(it)}\right)}{\delta \boldsymbol{\beta}}$ and $\frac{\delta S\left(\widehat{\boldsymbol{\theta}}^{(it)}\right)}{\delta \boldsymbol{\theta_t}}$ are the corresponding blocks of the Hessian matrix. The elements of the score vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta_t}$ are:

$$S(\boldsymbol{\beta}) = \frac{\sum_i X_i'\{\ y_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i'}\}}{\sigma^2} + \alpha_i x_i(T_i) - \exp(\boldsymbol{\gamma'} \boldsymbol{a}_i) \int \int_0^{T_i} h_0(s) a\ \boldsymbol{x}_i(s) \exp[\alpha\{\boldsymbol{x_i'}(s)\boldsymbol{\beta} + \boldsymbol{z_i'}(s) b_i\}] \times f\ (\boldsymbol{b}_i|\ T_i, \delta_i, y_i; \theta) ds\ d\boldsymbol{b}_i \quad (18)$$

$$S(\gamma) = \sum_i \boldsymbol{a}_i \left[\ \delta_i - \exp(\gamma' \boldsymbol{a}_i) \int \int_0^{T_i} h_0(s) \exp\left[a\ \boldsymbol{x}_i(s)\boldsymbol{\beta} + \boldsymbol{z_i'}(s) b_i\right] f\ (\boldsymbol{b}_i|\ T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) \times ds\ d\boldsymbol{b}_i \right] \quad (19)$$

$$S\ (\alpha) = \sum_i \delta_i \{\boldsymbol{x}_i'(T_i)\boldsymbol{\beta} + z_i'\ (T_i)\widetilde{\boldsymbol{b}_\iota}\} \left[- \exp(\boldsymbol{\gamma'} \boldsymbol{a}_i) \int \int_0^{T_i} h_0(s) \exp\left[a\ \boldsymbol{x}_i(s)\boldsymbol{\beta} + \boldsymbol{z_i'}(s) b_i\right] \times f\ (\boldsymbol{b}_i|\ T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) ds\ d\boldsymbol{b}_i \right] \quad (20)$$

$$S\ (\theta_{h_o}) = \sum_i \delta_i \frac{\delta\left(T_i; \theta_{h_o}\right)}{\delta \theta'_{h_o}} - \exp(\boldsymbol{\gamma'} \boldsymbol{a}_i) \int \int_0^{T_i} \frac{\delta\left(T_i; \theta_{h_o}\right)}{\delta \theta'_{h_o}} \exp\left[a\ \boldsymbol{x}_i(s)\boldsymbol{\beta} + \boldsymbol{z_i'}(s) b_i\right] \times f\ (\boldsymbol{b}_i|\ T_i, \delta_i, \boldsymbol{y_i}; \boldsymbol{\theta}) ds\ d\boldsymbol{b}_i \quad (21)$$

E. Hypothesis Testing

A simultaneous test is conducted to determine the effect of the coefficients as a whole in a model [13]. H_0: $\theta_1 = \theta_2 = \cdots = \theta_p = 0$ and H_1: there is at least one $\theta_p \neq 0$. The likelihood ratio test statistic (G^2) is as follows.

$$G^2 = -2 \ln \left[\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right] \quad (22)$$

In the G-test following the Chi-Square distribution χ^2, H_0 will be rejected if the value of $G > \chi^2(k;\alpha)$ or $p - value < \alpha$, which means that the predictor variables together mean the response variable. Meanwhile, the partial test is $H_0: \boldsymbol{\theta}_j = 0$; $H_1: \boldsymbol{\theta}_j \neq 0$, $j = 1,2,\dots,p$ with the test statistics

$$W = \left[\frac{\widehat{\boldsymbol{\theta}}_j}{SE\widehat{\boldsymbol{\theta}}_j}\right]^2 \tag{23}$$

Where $\widehat{\boldsymbol{\theta}}_j$ denotes the estimator of $\boldsymbol{\theta}_j$ and $SE\widehat{\boldsymbol{\theta}}_j$ is the error of $\widehat{\boldsymbol{\theta}}_j$, while W is a test with degrees of freedom equal to one. H_0 is rejected if $W > \chi^2_{(1;\alpha)}$ or $p - value < \alpha$, concluding that the predictor variable affects the response variable.

III. METHODOLOGY

The data used in the study is secondary data on diabetes mellitus in East Java obtained from IFLS in 2014 (late October 2015 to the end of April 2015) with IFLS5. Measurements were taken two to four times per patient. This data is the result of a survey that has been conducted. The dependent variables used include Survival time ($\boldsymbol{W}$). Survival time is the output of survival, where the patient is taking therapy for diabetes mellitus until he is declared dead or stops therapy during the study. Units in research in months. Patient status ($\boldsymbol{D}$), which consists of categories 1: Uncensored (Diabetes mellitus patient died) and 0: Censored (Diabetes mellitus patient transferred to another treatment or died from other causes). The level of HbA1c ($\boldsymbol{y_{ij}}$) and glycated hemoglobin (HbA1c) is the output of longitudinal. This marker is often used to predict the risk of complications in patients with diabetes mellitus [21]. While the independent variables in this study include the time of measurement ($\boldsymbol{X_1}$), measurements were made two to four times in 2007 on IFLS4 and 2014 on IFLS5. Medical history ($\boldsymbol{X_2}$) whether the patient is taking diabetes-lowering drugs. Gender ($\boldsymbol{X_3}$), Comorbidities ($\boldsymbol{X_4}$), which are diseases experienced by patients other than diabetes mellitus, and Complications ($\boldsymbol{X_5}$), where the patient's condition after treatment has complications or not.

Joint modeling is used in this research and the R program from JM packages is used to analyze the data [22], [23]. The model consists of longitudinal analysis with a linear mixed model and a survival analysis with a Cox proportional hazard approach. The stages of data analysis are as follows:

1. The first stage is collecting data on diabetes mellitus patients from IFLS5 according to the required variables.
2. Data analysis was conducted to determine the variables' general description.
3. It establishes a basic joint modeling model specification from the longitudinal and survival sub-model.
   a. LMM Model
      - Define the dependent and fixed effect (explanatory variable)
   b. Proportional Hazard Cox Model
      - Define the event and identify covariates
      - Define the baseline hazard function
4. We are estimating parameters using Maximum Likelihood Estimation (MLE).
5. After getting the parameter values, the results of the parameter values are entered in the model specifications so that a model from the joint model is formed, with a longitudinal sub-model using a linear mixed model approach and a survival sub-model with a proportional hazard Cox model.
   c. LMM Model
      - Fitting the model
      - Examine fixed effect estimates and the significance
      - Asses model fit
   d. Proportional Hazard Cox Model
      - Fitting the model
      - Examine hazard ratios and the significance
      - Asses model fit
6. Hypothesis testing was performed on the model formed to find the independent variables that affect the dependent variable so that the factors that are thought to affect the healing of patients with diabetes mellitus are known. This study chose to test the hypothesis using the Wald test with a significance level of 5%. H_0 is the independent variables, namely measurement time, treatment history, gender, comorbidities, and complications that do not significantly affect the healing time of diabetes mellitus. H_1 is the independent variables, namely measurement time, treatment history, gender, comorbidities, and complications that significantly affect the healing time of diabetes mellitus.

IV. RESULTS AND DISCUSSIONS

This study took IFLS data consisting of 293 observations with a minimum age of 16 to 65 years and made measurements of two to four repetitions in each observation. Characteristics in HbA1c levels, with the amount of HbA1c < **7**% [24]. The features of the survival time of patients with diabetes mellitus are in Table 1.

Table 1 Characteristics of survival time

Patient Data	Gender	Total (%)
Censored	Male	64.5
	Female	0.0
Died	Male	35.5
	Female	0.0

Table 1 shows that in this study, based on the characteristics of the survival time, there were more censored data than patients who died. There is 35.5% censored data in this case. Censored data include incomplete data, where patients are not taking therapy anymore or patients die. It is known that 64.5% of the data was censored with 189 patients. Two to four measurements were made until the total number of observations was 890 from 293 patients. In general, the data used in this research has the characteristics provided in the Table 2.

Table 2 Characteristics of data

Descriptive Statistics	X_1	X_2	X_3	X_4	X_5
Mean	4.35	0.49	0.92	0.58	0.68
Standard Deviation	4.81	0.50	0.27	0.49	0.47
Minimum	0.00	0.00	0.00	0.00	0.00
Maximum	18.00	1.00	1.00	1.00	1.00

In this case joint modeling is a linear mixed effects model and a Cox proportional hazard model for survival [13]. In the formation of joint modeling, there are longitudinal and survival sub-models. The model will be formed, and previously discussed, the longitudinal and survival sub-model. The following is an estimation of joint modeling using maximum likelihood by adding Newton Raphson's algorithm in diabetic patients.

Table 3 Estimates on data on diabetes mellitus patients

Parameter	Estimation	SE	*t-value*	*p-value*	Parameter
β_0	2.6401	0.0228	115.5422	0.0001*	β_0
β_1	-0.0017	0.0039	-4.4441	0.6570	β_1
β_2	0.0004	0.0048	0.0735	0.9414	β_2
σ_{00}	0.2821				σ_{00}
σ_{01}	0.0041				σ_{01}
γ_0	-0.2822	0.2008	-1.4053	0.1599	γ_0
α	1.0307	0.5734	1.7977	0.0722	α
γ_1	-8.7184	1.7723	-4.9193	0.0001*	γ_1
γ_2	-3.6026	1.7741	-2.0306	0.0423	γ_2
γ_3	-9.1032	1.7706	-5.1413	0.0001*	γ_3
γ_4	-4.4602	1.5829	-2.8177	0.0048	γ_4

A linear mixed effects model with random intercept is used to analyze longitudinal results. This model uses a variable $\boldsymbol{X_1}$, the measurement time, and $\boldsymbol{X_2}$, with a treatment history. This model shows that HbA1c levels in patients have progressed at different measurement times in the treatment history. To normalize HbA1c levels, use the square root of the amount of HbA1c in the longitudinal model. Analysis of longitudinal results using linear mixed effects and random slope for the square root of the total HbA1c levels. The longitudinal sub-model, where $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{\beta_1}$ is the parameter estimate of $\boldsymbol{X_1}$, and $\boldsymbol{\beta_2}$ is the parameter estimate of $\boldsymbol{X_2}$. $\boldsymbol{y_{ij}}$ shows as the square root of $\boldsymbol{j}$ the measurement of the number of HbA1c levels in individuals, with 293 individuals, with $\boldsymbol{b_i} = (\boldsymbol{b_{i0}}, \boldsymbol{b_{i1}})$ and $\boldsymbol{\sigma_{00}}$ being the standard deviation (individual), $\boldsymbol{\sigma_{01}}$ is the standard deviation at the time of measurement. In comparison, $\boldsymbol{\sigma^2}$ is a measurement of error.

The specifications in this model add random effects to the measurement time due to patients who do not take measurements at a predetermined time, resulting in missing data in the measurement of random effects. The added variable is measurement time. The parameters here will be estimated using the Restricted Maximum Likelihood (REML). REML is an estimation method used to estimate the fixed effect and variance parameters in a linear mixed model. The indicators used in the longitudinal sub-model are medication history and the relationship between medication history and measurement time. Based on Table 3 of the parameter estimation results, the form of the longitudinal sub-model is:

$$y_{ij} = 2{,}6403 - 0{,}0017t_{ij} + 0{,}0004\mathbf{X_2} \times t_{ij} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}$$

$$\boldsymbol{b_i} \sim \mathcal{N}(0, G)$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

with, the value of $\mathbf{cov}(\boldsymbol{G}) = \begin{pmatrix} \mathbf{0,2821} & \mathbf{0,0041} \\ \mathbf{0,0041} & \mathbf{-0,6756} \end{pmatrix}$ and the value of $\boldsymbol{\sigma^2} = \mathbf{0,4077}$.

After the model is known, the next step is to test the hypothesis. The model will be tested with a partial test by showing that the $\boldsymbol{p-value}$ of each independent variable is $\mathbf{0.0657}$ and $\mathbf{0.9414}$, respectively, where the value is more than $\boldsymbol{\alpha} = \mathbf{0,05}$. It can be concluded that the measurement time variables and treatment history have no significant effects. So that a significant longitudinal sub-model can be written:

$$y_{ij} = 2{,}6403 - \quad b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij}$$

because it is not significant, it will be continued on the joint modeling.

In the survival sub-model, we use the Cox proportional hazard model depending on the time of the study, namely the time the patient is taking treatment and the time he is declared cured or has stopped treatment. By entering covariates of patient data censored or died, treatment time ($\boldsymbol{X_1}$), history of treatment with ($\boldsymbol{X_2}$), gender ($\boldsymbol{X_3}$), comorbidities ($\boldsymbol{X_4}$), and complications ($\boldsymbol{X_5}$), with parameters and, where $\boldsymbol{\gamma_0}$ is a parameter of patients who did not take treatment within the specified study time. Meanwhile, an association measures the relationship between $\boldsymbol{m_i(t)}$ on the actual square root of the HbA1c level and the mortality factor. The parameter $\boldsymbol{\gamma}$ was applied with Cox proportional hazard regression. Based on Table 3, which is the result of the estimated parameters of the survival sub-model form as follows:

$$h_i(t|\mathcal{M}_i(t), X_i) = h_0(t)\exp\{-0{,}2822 - 8{,}7184X_{1i} - 3{,}6026X_{2i} - 9{,}1032X_{3i} - 4{,}4602X_{4i} - 7{,}1528X_{5i} + 1{,}0307\big(2{,}6403 - 0{,}0017t_{ij} + 0{,}0004X_2 \times t_{ij} + b_{0i} + b_{1i}t_{ij}\big)\}$$

After the model is known, the next step is to test the hypothesis. The model will be stretched with a partial test by showing that the $\boldsymbol{p-value}$ of each independent variable successively is $\mathbf{0.0001}$ where the value is less than $\boldsymbol{\alpha} = \mathbf{0,05}$. It can be concluded that the variables of measurement time, gender, and complications have a significant effect. So, the significant survival sub-model can be written as follows:

$$h_i(t|\mathcal{M}_i(t), X_i) = h_0(t)\exp\{-0{,}2822 - 8{,}7184X_{1i} - 9{,}1032X_{3i} - 7{,}1528X_{5i} + 1{,}0307\big(2{,}6403 - 0{,}0017t_{ij} + 0{,}0004\mathrm{X}_2 \times t_{ij} + b_{0i} + b_{1i}t_{ij}\big)$$

A joint model will be developed after the longitudinal and survival sub-model is constructed. From the longitudinal sub-model and survival sub-model, a joint model can be formed so that the joint model can be written as follows:

$$\begin{cases} h_i(t|\mathcal{M}_i(t), X_i) = h_0(t)\exp\{-0{,}2822 - 8{,}7184X_{1i} - \\ (-3{,}6026)X_{2i} - 9{,}1032X_{3i} - 4{,}4602X_{4i} - 7{,}1528X_{5i} + \\ 1{,}0307(2{,}6403 - \ 0{,}0017t_{ij} + 0{,}0004\mathrm{X}_2 \times t_{ij} + b_{0i} + b_{1i}t_{ij})\} \\ y_{ij} = 2{,}6403 - 0{,}0017t_{ij} + 0{,}0004\mathrm{X}_2 \times t_{ij} + b_{0i} + b_{1i}t_{ij} + \varepsilon_{ij} \\ b_i \sim \mathcal{N}(0, G), \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

with, the value of $\mathbf{cov}(\boldsymbol{G}) = \begin{pmatrix} \mathbf{0,2821} & \mathbf{0,0041} \\ \mathbf{0,0041} & \mathbf{-0.6756} \end{pmatrix}$ and the value of $\boldsymbol{\sigma^2} = \mathbf{0,4077}$.

After the model is known, the next step will be to test the hypothesis. The first is the simultaneous test, where the results of the joint modeling test can be carried out using the Wald test. The value of the test statistic was $\boldsymbol{W} = \mathbf{1081,872} \geq \boldsymbol{\chi_{\alpha;p2}} = \mathbf{11,07}$ and $\boldsymbol{p-value} = \mathbf{0,03175} < \mathbf{0,05}$. Because $\boldsymbol{W} \geq \boldsymbol{\chi_{\alpha;p2}}$ and $\boldsymbol{p-value} < \boldsymbol{\alpha}$, it is concluded that $\boldsymbol{H_0}$ is rejected. So simultaneously, there is at least one of independent variables that significantly affect the healing time of diabetes mellitus.

Based on the partial joint modeling test in Table 3 shows that the measurement of times ($\boldsymbol{X_1}$) and gender ($\boldsymbol{X_3}$) partially have a significant effect on diabetes mellitus. It is shown that the $\boldsymbol{p-value}$ of the independent variable is $\mathbf{0.0001}$, $\mathbf{0.0001}$, and $\mathbf{0.0001}$, respectively, where the value is less than $\boldsymbol{\alpha} = \mathbf{0.05}$, in contrast to previous research carried out by [7], namely obtaining two models with model 1 and model 2. In model 1, it is known that significant variables are measurement time and consumption of diabetes-lowering drugs with $\boldsymbol{p-values}$ of $\mathbf{0.000}$ and $\mathbf{0.0006}$, respectively, and the value of $\mathbf{0.05}$. In model 2, the significant variable is the measurement time with a $\boldsymbol{p-value}$ of 0.0000 and $\boldsymbol{\alpha} = \mathbf{0,05}$. So, the $\boldsymbol{p-value} < \boldsymbol{\alpha}$.

The two models were compared through the criteria for the best model using AIC. From of the AIC model 2 results, it is better because the value is smaller than model 1. Thus, a significant variable was obtained in previous study conducted by [25], [26] using longitudinal data: the measurement time. While in this study using the joint modeling method with estimates using the maximum likelihood and adding the Newton Raphson algorithm significant variables, namely the time of measurement of gender and complications.

Based on Table 3, the interpretation of the coefficients obtained in the formation of joint modeling states that the chance of recovering patients suffering from diabetes mellitus is influenced by the measurement time with a $\mathbf{-8.718}$ gender coefficient value of $\mathbf{-9.1032}$. At the same time, complications show a coefficient value of $\mathbf{-7.1528}$. Furthermore, the parameter shows the relationship between HbA1c levels and mortality factors affecting patient recovery with a coefficient of $\mathbf{1.0307}$. So, with the reduced measurement time, the patient has a lower possibility of recovering $\mathbf{8.7184}$ times. The variables of gender and complications also have a lower possibility of recovery, respectively, $\mathbf{9.1032}$ times and $\mathbf{7.1528}$ times.

This study provides the results of modeling and estimating diabetes mellitus data in Indonesia based on the method used. The study found that gender and complications significantly affect diabetes mellitus. One plausible explanation for this finding is that diabetes mellitus is more likely to occur in women [27], [28]. Other research shows that complications

greatly influence a person with diabetes mellitus [29], [30]. The study also discovered that separate modeling techniques employed to model and estimate diabetes mellitus in Indonesia produced inconsistent analytical outcomes, while joint model of the suggested data produced more accurate findings and developments. Additionally, we discovered that joint modeling, based on the analysis results, delivers more complicated information. The analysis employed in data analysis is based on statistical models. Making decisions can benefit from this. Therefore, policymakers, particularly the government, must use statistics and statistical modeling to inform their decisions. Additionally, researchers think providing the impact of joining two models impacts chosen for this study's techniques and models is crucial.

V. CONCLUSIONS AND SUGGESTIONS

Based on the results and discussion of parameter estimation using combined modeling analysis. The model is formed using longitudinal and survival sub-models. The result of the survival sub-model is that the explanatory variables have no significant effect. So, proceed to combined modeling. The combined modeling results show that the factors that have a significant effect on diabetes mellitus disease are time measurement, gender, and complications. The time measurement of patients who have a smaller chance of recovery is 8.718 times. The male gender of patients who have a smaller chance of recovery is 9.1032 times compared to women. As for complications, the coefficient value of patients who have a smaller chance of recovery is 7.1528 times. The interpretation of the model is that every time the time measurement decreases, the patient has a smaller chance of recovery, which is one time. The gender variable has a chance of recovery equal to time. For future research, the combined modeling results can be applied to other data. Combined modeling analysis can also be developed by considering right or left censored data. This study used the Cox proportional hazard method to accommodate survival time. Future research is expected to use other methods with different approaches and different algorithms.

REFERENCES

[1] S. Preethikaa and M. P. Brundha, "Awareness of diabetes mellitus among general population," *Res J Pharm Technol*, vol. 11, no. 5, pp. 1825–1829, 2018.

[2] I. D. Karantas, M. E. Okur, N. Ü. Okur, and P. I. Siafaka, "Dyslipidemia management in 2020: an update on diagnosis and therapeutic perspectives," *Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)*, vol. 21, no. 5, pp. 815–834, 2021.

[3] World Health Organization, "Global Report on Diabetes," 2016.

[4] D. Ratnawati, C. T. Wahyudi, and G. Zetira, "Dukungan Keluarga Berpengaruh Kualitas Hidup Pada Lansia dengan Diagnosa Diabetes Melitus," *Jurnal Ilmiah Ilmu Keperawatan Indonesia*, vol. 9, no. 02, pp. 585–593, 2019.

[5] S. Li *et al.*, "Prevalence of diabetes mellitus and impaired fasting glucose, associated with risk factors in rural Kazakh adults in Xinjiang, China," *Int J Environ Res Public Health*, vol. 12, no. 1, pp. 554–565, 2015.

[6] F. Z. Kamilah *et al.*, "Analysis of the determinants of diabetes mellitus in Indonesia: a case study of the 2014 indonesian family life survey," *Disease Prevention and Public Health Journal*, vol. 15, no. 2, p. 88, 2021.

[7] M. Wirastuti, "Analisis Longitudinal pada Data Pasien Diabetes Melitus," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 12, no. 1, pp. 13–19, 2019.

[8] D. M. Istuti, "Analisis Ketahanan Hidup Data Ties Pasien Tuberkulosis dengan Metode Exact Likelihood pada Model Regresi Cox Proportional Hazard," *Mathunesa: Jurnal Ilmiah Matematika*, vol. 7, no. 2, 2019.

[9] J. Tanoey and H. Becher, "Diabetes prevalence and risk factors of early-onset adult diabetes: results from the Indonesian family life survey," *Glob Health Action*, vol. 14, no. 1, p. 2001144, 2021.

[10] A. Saadane, E. M. Lessieur, Y. Du, H. Liu, and T. S. Kern, "Successful induction of diabetes in mice demonstrates no gender difference in development of early diabetic retinopathy," *PLoS One*, vol. 15, no. 9, p. e0238727, 2020.

[11] S. Schlesinger *et al.*, "Prediabetes and risk of mortality, diabetes-related complications and comorbidities: umbrella review of meta-analyses of prospective studies," *Diabetologia*, pp. 1–11, 2022.

[12] J. B. Cole and J. C. Florez, "Genetics of diabetes mellitus and diabetes complications," *Nat Rev Nephrol*, vol. 16, no. 7, pp. 377–390, 2020.

[13] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press, 2012.

[14] G. Ambi Ramakrishnan, K. K. Srinivasan, A. Mondal, and C. R. Bhat, "Joint model of sustainable mode choice for commute, shift potential and alternative mode chosen," *Transp Res Rec*, vol. 2675, no. 7, pp. 377–391, 2021.

[15] P. K. Mondal, "Joint modeling of longitudinal measurements and survival data with competing risks: application to HIV/AIDS study," University of Saskatchewan, Saskatoon, Canada, 2017.

[16] T. Yu, " Joint modelling of complex longitudinal and survival data, with applications to HIV studies," Thesis, Vancouver, 2019.

[17] A. Sattar and S. K. Sinha, "Joint modeling of longitudinal and survival data with a covariate subject to a limit of detection," *Stat Methods Med Res*, vol. 28, no. 2, pp. 486–502, 2019.

[18] P. Mondal, H. J. Lim, and O. C. S. Team, "The Effect of MSM and CD4+ Count on the Development of Cancer AIDS (AIDS-defining Cancer) and Non-cancer AIDS in the HAART Era," *Curr HIV Res*, vol. 16, no. 4, pp. 288–296, 2018.

[19] H. G. Vuong *et al.*, "BRAF mutation is associated with an improved survival in glioma—a systematic review and meta-analysis," *Mol Neurobiol*, vol. 55, pp. 3718–3724, 2018.
[20] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: With applications in R*. 2012. doi: 10.1201/b12208.
[21] T. Eriskawati, "KORELASI ANTARA KADAR HbA1c DAN RASIO LDL/HDL KOLESTEROL PADA PENDERITA DIABETES MELITUS TIPE 2," UNIVERSITAS SEBELAS MARET, Indonesia, 2015.
[22] R. C. Team, "A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria," *Available online: www. R-project. org/(accessed on 11 September 2020)*, 2018.
[23] D. Rizopoulos, "JM: An R package for the joint modelling of longitudinal and time-to-event data," *J Stat Softw*, vol. 35, pp. 1–33, 2010.
[24] L. F. Hidayati, "HUBUNGAN KADAR GULA DARAH DENGAN DERAJAT KEGOYAHAN GIGI PADA PASIEN DIABETES MELLITUS," Diploma, Poltekkes Kemenkes Yogyakarta, Yogyakarta, Indonesia, 2019.
[25] S. T. Certo, M. C. Withers, and M. Semadeni, "A tale of two effects: Using longitudinal data to compare within- and between-firm effects," *Strategic Management Journal*, vol. 38, no. 7, pp. 1536–1556, 2017.
[26] M. Herle *et al.*, "Identifying typical trajectories in longitudinal data: modelling strategies and interpretations," *Eur J Epidemiol*, vol. 35, no. 3, pp. 205–222, 2020.
[27] C. Deischinger *et al.*, "Diabetes mellitus is associated with a higher risk for major depressive disorder in women than in men," *BMJ Open Diabetes Res Care*, vol. 8, no. 1, p. e001430, 2020.
[28] Z. Gao, Z. Chen, A. Sun, and X. Deng, "Gender differences in cardiovascular disease," *Med Nov Technol Devices*, vol. 4, p. 100025, 2019.
[29] R. Balaji, R. Duraisamy, and M. P. Kumar, "Complications of diabetes mellitus: A review.," *Drug Invention Today*, vol. 12, no. 1, 2019.
[30] M. J. L. Verhulst, B. G. Loos, V. E. A. Gerdes, and W. J. Teeuw, "Evaluating all potential oral complications of diabetes mellitus," *Front Endocrinol (Lausanne)*, vol. 10, p. 56, 2019.