EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol7.iss1.art1

Multi-state Models for Longitudinal Data with Hidden Markov Method

Yadhurani Dewi Amritha[1*], Danardono[2]

[1] Information Technology Study Program, Faculty of Engineering and Informatics, Universitas Pendidikan Nasional
[2] Actuarial Study Program, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada

[*] *Corresponding author: yadhurani@undiknas.ac.id*

Abstract: Monitoring disease progression in chronic conditions involves uncertainty due to irregular observation intervals. This study employs a multi-state Hidden Markov Model (HMM) to analyze longitudinal data on Type 2 Diabetes Mellitus (T2DM) obtained from BPJS Kesehatan (220 patients, 2,855 visits). HMM was selected for its robust capacity to manage irregular observation schedules and estimate "hidden" health states not directly observed in clinical records. We estimated transition intensities and probabilities using the maximum likelihood method across four model specifications. Based on the Akaike Information Criterion (AIC), the model incorporating gender covariates in all transitions demonstrated the best fit (AIC = 3212), significantly outperforming the baseline model (AIC = 3235). Quantitative findings reveal that gender significantly influences progression; specifically, male patients have a 1.87 times higher risk of progressing from single to multiple complications (Hazard Ratio = 1.87, 95% CI: 1.02–3.46) compared to females. This study concludes that discrete-time HMM is a highly effective tool for extracting insights from administrative health data, suggesting the need for gender-differentiated chronic care strategies in Indonesia.

Keywords: Hidden markov model, Longitudinal data, Type 2 diabetes mellitus, Transition probability

Introduction

An individual's health is not static and can transition between conditions as time progresses. The multi-state model framework provides a method to quantitatively examine such changes. For instance, this model can be applied to cases of diabetes mellitus, where a patient may move through various phases of the disease. Starting with a state without any complications, then there is one complication, multiple complications until death occurs. In this change, it takes a long time before an event occurs, so the data to be obtained will be survival data. In survival analysis, the time when observation begins must be determined in such a way that each individual begins to be observed at the same point. It is also important to determine the exact end point of the survival analysis so the time is well-defined.

Multi-state Markov models provide a framework for describing the trajectory of chronic conditions like cancer, stroke, and diabetes mellitus. These models are defined by the Markov property, a 'memoryless' principle. This means that the likelihood of any

future transition depends solely on the current state, not on the sequence of states that preceded it. Markov multi-state models are widely used due to their simplicity and the theoretical convenience of the memoryless property. However, these standard models rely on a critical assumption: that the true state of the process is perfectly observed at each time point. In the progression of Type 2 Diabetes Mellitus, this assumption is often violated. Patients may transition between underlying biological states—for instance, developing an asymptomatic complication—without immediate clinical detection. This creates a discrepancy where the 'observed' state in the medical record differs from the patient's 'true' physiological state. Because standard Markov models cannot account for this classification error, applying them to such data would yield biased estimates. Consequently, the standard framework is insufficient, necessitating an extension that can distinguish between the latent disease process and the observed clinical data.

The Hidden Markov Model (HMM) is an extension of the traditional Markov Model. The key distinction in an HMM is that the states are not directly observed; instead, each state is associated with a probability distribution that governs the observations. This model can then be applied to cases where the process cannot be observed directly (hidden) but only be observed through a collection of other stochastic processes that produce observation stages as much as possible [1]. The suitability of the HMM framework for analyzing longitudinal data stems from its capacity to define the relationship between two stochastic processes: a manifest outcome process and a latent state process. Consequently, these models, which rely on an underlying discrete-time Markov chain, are prevalent in various domains, notably sound and signal processing [2] and for modeling biological sequence data [3].

The Hidden Markov Model (HMM) has demonstrated remarkable versatility across various disciplines. Beyond biological applications, its predictive power has been proven in diverse fields such as stock market forecasting, as described by Zulfikar et al. [5]. In the medical domain, HMM serves as a robust tool for monitoring complex health conditions. Jackson et al. [4] provided a foundational framework for this approach by applying HMM to heart transplant data to handle classification errors in disease progression. Similarly, application-driven studies, such as Wicaksana et al. [6], have utilized HMM for tracking women's fertility. Furthermore, methodological advancements continue to evolve, with Zhou et al. [7] proposing continuous-time frameworks for longitudinal data. Building on these successful applications, this study adopts the HMM framework specifically to address the challenges of irregular observation intervals in Type 2 Diabetes Mellitus data.

Despite the established utility of multi-state models, a critical research gap remains in their application to real-world administrative health datasets, particularly in developing nations like Indonesia. Most existing studies rely on clinical trial data with fixed observation schedules, which fail to reflect the complexity of administrative claims data characterized by irregular observation intervals and potential misclassification of disease states. Consequently, the progression of chronic diseases in such databases remains poorly understood.

To address this gap, this study aims to develop and apply a multi-state HMM to analyze longitudinal data on Type 2 Diabetes Mellitus. This approach is specifically chosen to evaluate the impact of significant covariates on survival model functions within the context of the BPJS Kesehatan dataset. By utilizing the HMM framework, this research overcomes the limitations of traditional models by connecting a series of observable outcomes (clinical diagnoses) with an unobservable underlying process, thereby providing a more accurate representation of the true, latent states of disease severity.

The specific methodology employed is a multi-state HMM, which is chosen for its superior flexibility in handling the complexities of real-world clinical data. Unlike traditional models that may require uniform observation schedules, the multi-state HMM can effectively accommodate longitudinal data collected at regular, irregular, or even continuous intervals. This capability is crucial for analyzing administrative health datasets where patient visits are not systematically scheduled, thus ensuring that the integrity of the data is maintained without resorting to problematic imputation methods.

For the estimation of model parameters, this research will utilize the robust Maximum Likelihood Estimation (MLE) method. This statistical technique identifies the parameter values that are most likely to have produced the observed data. Furthermore, to ensure the development of a model that is both accurate and parsimonious, a rigorous model selection process will be conducted. Various model specifications, both with and without covariates, will be compared using the Akaike Information Criterion (AIC). This approach allows for a balanced assessment of model fit against model complexity, leading to an efficient and reliable estimation procedure.

The primary contribution of this research is the establishment of a flexible and statistically sound framework for managing and interpreting complex longitudinal data. The practical utility and validity of this model will be demonstrated through an in-depth case study on the progression of Type 2 Diabetes Mellitus, a classic example of a chronic, multi-stage illness. The analysis will be conducted using a comprehensive dataset from Indonesia's Social Security Agency on Health (BPJS Kesehatan) at the Advanced Referral Health Facilities (FKRTL) level, grounding the research in a relevant, real-world public health context.

Materials and Methods

Materials

This chapter contains the list of materials and tools used in this research. The names of materials and tools used are written in detail and accompanied by units and brands.

- Data Source: Secondary data from the Social Security Agency on Health of Indonesia (BPJS Kesehatan) for the period of January 2015 to December 2020.
- Data Structure and Variables: The dataset consists of longitudinal records of Type 2 Diabetes Mellitus (DM) covering 220 patients with a total of 2,855 clinical visits. The data is structured in a long format, where each row

represents a distinct observation time point for a specific patient. The key variables used in the analysis are

- Patient ID: Anonymized unique identifier for each subject.
- Visit Time: The time of observation measured in months.
- Patient Status: The clinical condition of the patient at each visit (State 1–4).
- Gender: A binary demographic covariate.

- Software: R software (Version 4.13).
- Analysis Package: The msm library (Version 0.7) in R.

Methods

This section thoroughly details the research methodology, from a comprehensive description of the materials and instruments used to a transparent account of the procedural steps. The procedures for data acquisition, processing, and interpretation are outlined in a clear and systematic manner.

This research employed a quantitative analysis of secondary data. The analytical procedure was performed as follows:

1. Data Collection and Selection: The study utilized longitudinal data of Type 2 Diabetes Mellitus patients from BPJS Kesehatan (2015-2020). A sample of 220 patients was selected from this dataset. The data was treated as multi-state data, where a patient's condition is assumed to transition between different states over time.
2. Statistical Modeling: The analysis was conducted using a Hidden Markov Model (HMM). While traditional approaches like Cox Proportional Hazards models are effective for analyzing the time to a single terminal event, they are less suitable for chronic diseases involving multiple reversible transitions between disease stages. Similarly, linear mixed models are typically used for continuous longitudinal outcomes rather than discrete state transitions. In contrast, HMM is explicitly chosen for this study because it addresses a critical limitation in administrative health data: classification uncertainty. In real-world clinical data, a patient's recorded diagnosis (observed state) may not always perfectly reflect their true physiological condition (hidden state) due to screening limitations or recording errors. HMM allows us to model the true latent disease progression while accounting for probabilistic errors in the observed data, as illustrated in the conceptual diagram in Figure 1

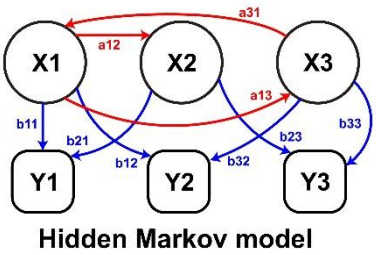

Figure 1. Conceptual Diagram of the Hidden Markov Model

3. Model Parameter Estimation: The analysis involved estimating several key parameters:

- Transition Intensity: This represents the instantaneous rate at which a patient transitions from one health state to another. The effect of covariates, such as gender, on these intensities was modeled using a proportional intensity model:

$$q_{ij}(z(t)) = q_{ij}^{(0)} \exp(\beta_{ij}^{T} z(t)) \quad (1)$$

- Transition Matrix (Q) : The transition matrix was estimated via the mean sojourn time, which is the average time spent in a state before moving to another. For a constant intensity q_{ii}, the mean sojourn time is calculated as:

$$E(T_i) = -\frac{1}{q_{ii}} \quad (2)$$

- Likelihood Function: Model parameters, specifically $q_{ij}^{(0)}$ and β_{ij} , were estimated by maximizing the likelihood function based on the observed data.

4. Technical Implementation: All statistical modeling and analyses were performed using R software version 4.13, specifically utilizing the msm package version 0.7, which is designed for multi-state modeling.

Results and Discussions

Hidden Markov Model

The model to be used in this research is a multi-state model with 4 states. In this study, the disease progression is categorized into four discrete health states (S_t) representing the patient's condition at time t:

State 1: Type 2 DM without complications.

State 2: Type 2 DM with a single complication.

State 3: Type 2 DM with multiple complications.

State 4: Died (Absorbing state).

The Hidden Markov Model (HMM) was explicitly chosen for this study over traditional survival analysis methods due to the complex nature of diabetes progression. The selection of HMM is justified by three key advantages over alternative approaches:

- Versus Cox Models: While Cox Proportional Hazards models are effective for analyzing the time to a single terminal event (e.g., survival time), they are ill-suited for chronic diseases involving multiple, reversible transitions. HMMs naturally handle these complex dynamics, such as a patient improving from 'State 2' back to 'State 1', which traditional survival models cannot easily accommodate.
- Versus Mixed Models: Unlike Linear Mixed Models, which are typically designed for continuous longitudinal outcomes (e.g., blood glucose levels), HMMs are specifically tailored for categorical state data (e.g., discrete disease stages).
- Handling Measurement Error: Most critically, HMM accounts for the distinction between the "true" underlying health status (hidden state) and the recorded clinical diagnosis (observed state). This capability is essential for administrative datasets like BPJS, where classification errors or screening limitations may occur. By modeling the probability of error, HMM provides a more accurate estimation of disease progression than models that assume observed data is perfect.

The conceptual framework illustrating the relationship between the hidden physiological process and the observed data is presented in Figure 1.

The unit of time that used is month. Information about the patient's healthy time is not available in the data because the patient only visits when they are sick. The model operates under the Hidden Markov assumption, which governs the progression of a patient through different health states. Furthermore, the framework permits non-adjacent transitions between states, a possibility that is visualized in Figure 2.

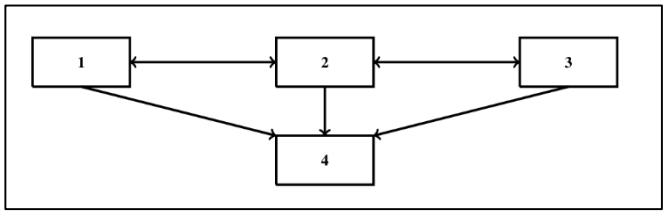

Figure 2. Multi-state model

From the state transition diagram shown in Figure 1, the transition intensity matrix (Q) can be constructed as follows:

$$Q = \begin{pmatrix} q_{11} & p_{12} & 0 & q_{14} \\ q_{21} & q_{22} & q_{23} & p_{24} \\ 0 & q_{32} & q_{33} & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (3)$$

With $q_{ij} = -\sum_{j \neq i} q_{ij}$ and $i, j = 1,2,3,4$

Models without covariates can be modeled for the first time and can be adjustable by the transition intensity matrix Q. The model is built on the assumption that a patient may transition to an adjacent state—representing either progression or recovery—during the course of the illness. It further stipulates that death is an absorbing state that can be entered from any of the living states (State 1, State 2, or State 3). By using Equation 2 and using the msm library version 0.7 in R software version 4.13, the value of Q is obtained, which is an estimate for the transition intensity as in Equation 4.

$$Q = \begin{pmatrix} -0.05 & 0.025 & 0 & 0.025 \\ 0.0167 & -0.05 & 0.0167 & 0.0167 \\ 0 & 0.025 & -0.05 & 0.025 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad (4)$$

The Q matrix, which specifies the allowed transitions between states, was passed as the qmatrix argument to the msm function. The parameter estimates generated by this model are presented in Table 1.

Table 1. Result of msm command

State	Baseline	
State 1 – state 1	-0.196458	(-0.232648,-0.16590)
State 1 – state 2	0.191305	(0.160916, 0.22743)
State 1 – state 4	0.005154	(0.002221, 0.01196)
State 2 – state 1	0.145886	(0.122022, 0.17442)
State 2 – state 2	-0.191470	(-0.222645,-0.16466)
State 2 – state 3	0.040697	(0.030606, 0.05412)
State 2 – state 4	0.004887	(0.002193, 0.01089)
State 3 – state 2	0.164159	(0.124328, 0.21675)
State 3 – state 3	-0.170167	(-0.222986,-0.12986)
State 3 – state 4	0.006008	(0.001369, 0.02636)
2 * log-likelihood: 3221.347		

With the estimated transition probability in a period of one month for model 1 is as follows:

$$P = \begin{pmatrix} 0.83315336 & 0.1584865 & 0.003241293 & 0.005118838 \\ 0.12085929 & 0.8400671 & 0.034160713 & 0.004912866 \\ 0.00997037 & 0.1377946 & 0.846329523 & 0.005905538 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (5)$$

So that the value of $P(t = 10)$ obtained as follows:

$$P(t = 10) = \begin{pmatrix} 0.40 & 0.47 & 0.08 & 0.050 \\ 0.36 & 0.49 & 0.11 & 0.049 \\ 0.25 & 0.44 & 0.26 & 0.054 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (6)$$

Models with covariates can be formed into three models, namely model with covariates present in all transitions, model with transition-specific covariates and model with constrained covariate effects. For the first model that has a covariate in each transition, it is called model 2. The results of the first model with covariates are as follows:

Table 2. Model Results with Covariates

State	Baseline		Gender	
State 1 – state 1	-0.1952096	(-2.403e-01, -1.586e-01)		
State 1 – state 2	0.1950970	(1.636e-01, 2.326e-01)	1.371e+00	(9.387e-01, 2.002e+00)
State 1 – state 4	0.0001126	(4.241e-88, 2.991e+79)	1.069e+05	(2.589e-228,4.417e+237)
State 2 – state 1	0.1470881	(1.223e-01, 1.769e-01)	2.829e+00	(1.891e+00, 4.233e+00)
State 2 – state 2	-0.1921991	(-2.247e-01, -1.644e-01)		
State 2 – state 3	0.0409599	(3.074e-02, 5.457e-02)	1.874e+00	(1.016e+00, 3.457e+00)
State 2 – state 4	0.0041510	(1.463e-03, 1.178e-02)	4.148e-01	(7.465e-02, 2.305e+00)
State 3 – state 2	0.1632000	(1.234e-01, 2.158e-01)	1.197e+00	(6.596e-01, 2.173e+00)
State 3 – state 3	-0.1633012	(-2.349e-01, -1.135e-01)		
State 3 – state 4	0.0001013	(1.440e-167, 7.120e+158)	3.185e+05	(0.000e+00, Inf)
-2 * log-likelihood: 3184				

The estimated transition probability for model 2

$$P = \begin{pmatrix} 0.835 & 0.162 & 0.00334 & 0.000460 \\ 0.122 & 0.840 & 0.03449 & 0.003807 \\ 0.010 & 0.137 & 0.85216 & 0.000396 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (7)$$

The results that show on table 3 are a model that still needs to be simplified because the intensity matrix represents 83 male patients and 137 female patients. To extract the intensity matrix separately between men and women. So the results are as follows:

Table 3. Model Results with Separated Covariates Male Gender

State	State 1	State 2	State 3	State 4
State 1	-1.594e-01 (-2.183e-01, -1.164e-01)	1.594e-01 (1.164e-01, 2.183e-01)	0	6.700e-08 (1.623e-240, 2.766e+225)
State 2	7.549e-02 (5.375e-02, 1.060e-01)	-1.102e-01 (-1.444e-01, -8.407e-02)	2.738e-02 (1.653e-02, 4.534e-02)	7.299e-03 (3.780e-03, 1.409e-02)
State 3	0	1.454e-01 (8.902e-02, 2.375e-01)	-1.454e-01 (-2.375e-01, -8.902e-02)	2.992e-08 (0.000e+00, Inf)
State 4	0	0	0	0

Table 4. Model Results with Separated Covariates Female Gender

State	State 1	State 2	State 3	State 4
State 1	-0.225622 (-0.2770399,-0.18375)	0.218457 (0.1769773, 0.26966)	0	0.007165 (0.0033483, 0.01533)
State 2	0.213576 (0.1719770, 0.26524)	-0.267911 (-0.3229170,-0.22227)	0.051308 (0.0362604, 0.07260)	0.003028 (0.0006213, 0.01475)
State 3	0	0.174083 (0.1240859, 0.24422)	-0.183611 (-0.2540322,-0.13271)	0.009529 (0.0024571, 0.03695)
State 4	0	0	0	0

The hazard.msm function provides an estimate of the hazard ratio corresponding to each covariate effect on the transition intensity. The results of the hazard function are as follows:

Table 5. Hazard Ratio For Transition

State	HR	L	U
State 1 – State 2	1.37e+00	9.39e-01	2.00e+00
State 1 – State 4	1.07e+05	2.59e-228	4.42e+237
State 2 – State 1	2.83e+00	1.89e+00	4.23e+00
State 2 – State 3	1.87e+00	1.02e+00	3.46e+00
State 2 – State 4	4.15e-01	7.47e-02	2.30e+00
State 3 – State 2	1.20e+00	6.60e-01	2.17e+00
State 3 – State 4	3.19e+05	0.00e+00	Inf

Interpretation of Model 2 ResultsThe hazard ratios (HR) presented in Table 5 reveal critical insights into gender differences in disease progression. An HR greater than 1 indicates an increased rate of transition. Notably, the transition from 'Type 2 DM with 1 complication' to 'Multiple complications' (State 2 →State 3) has an HR of 1.87 (95% CI: 1.02–3.46). This implies that, in this cohort, male patients are approximately 1.87 times more likely to experience disease progression compared to females. Conversely, the transition from 'State 2' back to 'State 1' (Recovery) also shows a high HR of 2.83, suggesting a higher volatility in the male group—they are prone to both faster progression and faster recovery compared to the baseline female group.

Model 3 SpecificationFollowing the evaluation of the full covariate model, we tested a more parsimonious specification. The second covariate-based model, designated as Model 3, assumes that the gender effect is not universal across all transitions but is specific to certain pathways. Specifically, Model 3 constrains the covariate effects to only the onset pathways: 1→2 (progression to single complication) and 1→4 (mortality from baseline). The results for this transition-specific model are obtained as follows:

Table 6. Model Results With Transition-Specific Covariates

State	Baseline		Gender	
State 1 – State 1	-0.1962163	(-2.362e-01, -0.16297)		
State 1 – State 2	0.1953115	(1.642e-01, 0.23233)	0.8258	(6.010e-01,1.135e+00)
State 1 – State 4	0.0009048	(1.782e-10,4595.26781)	198.8024	(4.454e-17,8.873e+20)
State 2 – State 1	0.1465238	(1.225e-01, 0.17527)	1.0000	
State 2 – State 2	-0.1928518	(-2.243e-01, -0.16584)		
State 2 – State 3	0.0407409	(3.063e-02, 0.05418)	1.0000	
State 2 – State 4	0.0055871	(2.940e-03, 0.01062)	1.0000	
State 3 – State 2	0.1644516	(1.246e-01, 0.21708)	1.0000	
State 3 – State 3	-0.1706129	(-2.235e-01, -0.13023)		
State 3 – State 4	0.0061613	(1.376e-03, 0.02758)	1.0000	
-2 * log-likelihood: 3218				

The estimated transition probability for model 3

$$P = \begin{pmatrix} 0.834 & 0.162 & 0.00331 & 0.00131 \\ 0.121 & 0.839 & 0.03417 & 0.00528 \\ 0.010 & 0.138 & 0.84596 & 0.00608 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (8)$$

The third model with covariate that can be formed is a model with constrained covariate effects, it is called model 4. This model limits the effect of the gender covariate to be the same at the disease progression level. q_{12}, q_{23}, becomes the same at the death rate i.e. q_{14}, q_{24}, q_{34} and becomes the same for the recovery rate i.e. q_{21}, q_{32}. The intensity parameters are summed by reading all the rows of the transition matrix and starting from the first row: $(q_{12}, q_{14}, q_{21}, q_{23}, q_{24}, q_{32}, q_{34})$, so the boundary indicator becomes (1, 2, 3, 1, 2, 3, 2). Any increasing number vector can be used as an indicator. The results can be obtained as follows:

Table 7. Model Results with Constrained Covariate Effects

State	Baseline		Gender	
State 1 – State 1	-0.196715	(-0.233990,-0.16538)		
State 1 – State 2	0.191923	(0.160696, 0.22922)	1.500	(1.0912,2.061)
State 1 – State 4	0.004793	(0.001954, 0.01176)	1.379	(0.6338,2.999)
State 2 – State 1	0.149221	(0.124390, 0.17901)	2.241	(1.6027,3.132)
State 2 – State 2	-0.195082	(-0.227476,-0.16730)		
State 2 – State 3	0.040686	(0.030580, 0.05413)	1.500	(1.0912,2.061)
State 2 – State 4	0.005175	(0.002367, 0.01132)	1.379	(0.6338,2.999)
State 3 – State 2	0.152375	(0.114663, 0.20249)	2.241	(1.6027,3.132)
State 3 – State 3	-0.157683	(-0.208075,-0.11950)		
State 3 – State 4	0.005309	(0.001093, 0.02579)	1.379	(0.6338,2.999)
-2 * log-likelihood: 3197				

The estimated transition probability for model 4

$$P = \begin{pmatrix} 0.83323 & 0.159 & 0.00326 & 0.00481 \\ 0.12339 & 0.837 & 0.03431 & 0.00514 \\ 0.00949 & 0.128 & 0.85674 & 0.00528 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (9)$$

Model Assessment

In the case study of type 2 DM, 4 models can be formed, starting where there is no covariate effect, there is one model and three models where there is a covariate effect. To determine the optimal model specification, we compared the four candidate models using the Akaike Information Criterion (AIC), as summarized in Table 8. The AIC serves as a measure of the trade-off between the model's goodness of fit and its complexity; a lower AIC value indicates a better model.

Table 8. AIC value on the model

No	Model	AIC value
1	Model 1 (without covariates)	3235
2	Model 2 (with covariates on all transitions)	3212
3	Model 3 (with covariates at transitions 12 and 14)	3236
4	Model 4 (with covariates on finite transitions)	3217

As observed in Table 8, Model 2 (Full Covariates) yielded the lowest AIC value of 3212. This represents a substantial improvement over the baseline Model 1 (AIC = 3235), confirming that incorporating gender as a covariate significantly improves the model's explanatory power.

Furthermore, Model 2 also outperformed the restricted specifications (Model 3 and Model 4). This result implies that the effect of gender is not limited to specific pathways (as assumed in Model 3) nor is it uniform across transitions (as assumed in Model 4). Instead, gender appears to have a distinct and significant influence on every stage of the disease progression. Consequently, Model 2 is selected as the final model for interpreting the dynamics of Type 2 Diabetes Mellitus in this cohort.

Conclusion

Based on the discussion and case studies described in the previous chapter, the following conclusions can be written as follows:

a) This study utilizes a multi-state Hidden Markov Model (HMM) to analyze longitudinal data and estimate the influence of covariates on survival model functions. The HMM is particularly suitable for this task because it can model multi-state data where the true underlying status of a subject is unobservable (hidden), but can be inferred from a series of measurable indicators. Covariate effects can form a variety of models where these effects can affect each status or only certain status. Furthermore, the best model can be seen from the AIC value. Through AIC value, it can be determine the best model among others.

b) The results of a case study on type 2 diabetes mellitus patients for the Hidden Markov multi-state model are obtained 4 models, the best model is used gender covariates in each state which can be seen from the AIC value. The

AIC value that using covariates in all states have the smallest value if compared to other models.

In this research, there are many things to develop in the future, so there are several suggestions proposed for further research which are related to the hidden Markov model, such as the implementation of hidden Markov models other than biostatistics topics with more complex data and analysis can be implemented.

Acknowledgement

We would like to thank you to all entities that helped and encouraged me to complete this research. I would also like to thank and appreciate to BPJS Kesehatan for the sample data for this research.

References

[1] D. Rosadi, Pengantar Analisa Runtun Waktu. Yogyakarta: FMIPA Universitas Gadjah Mada, 2006.
[2] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," Technometrics, vol. 33, no. 3, pp. 251–272, 1991.
[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[4] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, "Multistate Markov models for disease progression with classification error," The Statistician, vol. 52, no. 2, pp. 193–209, 2003.
[5] M. Zulfikar, A. Bado, K. Jaya, M. R. Hassan, and B. Nath, "Implementasi Hidden Markov Model pada peramalan data saham," Program Studi Statistika FMIPA, Univ. [Nama Universitas], [Kota], 2020.
[6] I. S. Wicaksana, S. Setiawidayat, and D. U. Effendy, "Metode Hidden Markov Model untuk pemantauan masa subur wanita berbasis Android," J. Appl. Sci. Electr. Eng., vol. 1, no. 1, pp. 26–39, 2020.
[7] J. Zhou, K. Kang, T. Kwok, and X. Song, "Joint Hidden Markov Model for longitudinal and time-to-event data with latent variables," Multivariate Behav. Res., vol. 57, no. 3, pp. 441–457, 2022.
[8] Danardono, Analisis Data Longitudinal. Yogyakarta: Gadjah Mada Univ. Press, 2015.
[9] E. T. Lee and J. W. Wang, Statistical Methods for Survival Data Analysis, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2003.
[10] S. Karlin and H. M. Taylor, A First Course in Stochastic Processes, 2nd ed. New York, NY, USA: Academic Press, 1975.
[11] P. Dymarski, Hidden Markov Models: Theory and Applications. Rijeka, Croatia: InTech, 2011.
[12] American Diabetes Association, "Standards of medical care in diabetes—2007," Diabetes Care, vol. 30, no. suppl 1, pp. S4–S41, 2007.
[13] World Health Organization, "Global Report on Diabetes," World Health Organization, Geneva, Switzerland, 2016.

[14] H. Anton and C. Rorres, Elementary Linear Algebra: Applications Version, 10th ed. Hoboken, NJ, USA: John Wiley & Sons, 2011.
[15] C. H. Jackson and L. D. Sharples, "Multi-state Markov models for onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients," Stat. Med., vol. 21, no. 1, pp. 113–128, 2002.
[16] C. H. Jackson, "Multi-state models for panel data: The msm package for R," J. Stat. Softw., vol. 38, no. 8, pp. 1–28, 2011.
[17] X. Li, Y. Zhuang, B. Lu, and G. Chen, "A multi-stage Hidden Markov Model of customer repurchase motivation in online shopping," Decis. Support Syst., vol. 120, pp. 72–80, 2019.
[18] Kementerian Kesehatan RI, "Hasil Utama Riset Kesehatan Dasar (RISKESDAS) 2018," Kementerian Kesehatan Republik Indonesia, Jakarta, Indonesia, 2018.
[19] G. A. Satten and I. M. Longini Jr., "Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease," Appl. Stat., vol. 45, no. 3, pp. 275–309, 1996.
[20] J. Zhou, S. Song, and L. Sun, "Continuous time Hidden Markov Model for longitudinal data," J. Multivariate Anal., vol. 179, Art. no. 104640, 2020.