

A hidden semi-Markov model with explicit state duration for malaria symptoms in children below five years

Esther Obure Ogetange¹

Drinold Mbete²

Annette Okoth³

¹estherogetange@gmail.com

²dmbete@mmust.ac.ke

³aokoth@mmust.ac.ke

^{1,2,3}Department of Mathematics, Masinde Muliro University of Science and Technology, Kenya

<https://doi.org/10.51867/ajernet.7.4.84>

Abstract

Malaria disease is the primary cause of death and morbidity among children below five years. Globally, children below five years account for 76% of all malaria deaths; in Africa, which has the highest malaria disease burden worldwide, this percentage increases to 78.1%. In Kenya, 80% of all malaria cases are reported among children below five years. Despite detailed literature on malaria disease among children below five years, there is scarce literature on malaria symptoms modeling in children below five years and yet these symptoms are common. Therefore, this study formulated a Hidden Semi-Markov model (HSMM) for malaria symptoms in children below five years. A retrospective study design was used and data was obtained from health records of children below five years at the Vihiga County Teaching and Referral Hospital, Kenya, between the months of January to December 2023. The data comprised of 1281 observations from 300 children below five years, with 2 to 7 repeated observations per child, on fever, vomiting, cough, rigors, sweating, pallor, convulsion, prostration and diarrhoea symptoms. The model assumes that the next state depends only on the current state and that the time spent in a state follows a state-specific duration distribution, with no self-transitions. The Expectation-Maximization approach was used to estimate the model parameters. The findings revealed that 32.55% (49% of the children) of all observations were in the severe state, followed by 23.81% (21% of the children) healthy, 22.33% (16% of the children) moderate and 21.31% (16% of the children) mild. Additionally, state duration estimation showed that children tended to remain in the healthy, mild, and moderate states for an average of two days each (1.96, 1.76 and 1.80 days), while the severe state lasted for approximately three days (2.77 days). The findings, especially the short durations in earlier states, highlight how rapidly a child transitions to the severe state, where the risk of mortality is high. This underscores the need for early diagnosis and timely medical intervention to mitigate mortality risks in children. The study

Licensed Under Creative Commons Attribution (CC BY-NC)

is limited by its use of a single hospital, retrospective records, and possible classification errors; thus, prospective multi-hospital studies and external validation are recommended.

MSC2010 Subject Classification: 60J10, 62F10, 62P10

Keywords: Children under five, expectation-maximization, hidden semi-markov model, explicit state duration, malaria symptoms.

1 Introduction

Malaria is an infectious illness that has gained importance recently due to its frequent occurrence. It is spread by a female *Anopheles* mosquito carrying *Plasmodium* parasites (Li & Ma, 2022). Despite the many years of research and intervention, malaria continues to be the leading cause of morbidity and death among children below five years. The World Health Organization (WHO) confirms the vulnerability of this age group to exposure and burden of this disease through their latest malaria report of the year 2024 that says, malaria cases were estimated to be at 263 million cases globally with 597,000 deaths in the year 2023, out of which 76% accounts for deaths in children below five years (WHO, 2024). Africa accounted for 94% of all malaria cases worldwide, which is estimated to be 246 million in the year 2023, with 569,000 deaths. 78.1% of all deaths in Africa region were reported among children below five years. In Kenya, the disease caused 3.3 million cases and 11,478 deaths, 80% of all these cases being reported among children below five years (Lacey et al., 2023).

According to a study by Odhiambo (Odhiambo, 2023), in Kenya the highest malaria burden is concentrated in six counties; Busia, Kisumu, Kakamega, Migori, Siaya and Vihiga. Although these counties share the same climatic conditions conducive for malaria transmission, Vihiga County stands out due to its remarkably adjusted malaria incidence of 748 cases per 1,000 people, the highest in the region. A malaria data quality assessment by the Kenya Field Epidemiology Training Program (K-FELTP) shows that Vihiga County reported 156,746 confirmed malaria cases in the outpatient summary against 45,718 in the laboratory summary, for the period between August 2020 and July 2021 (Ouma, 2021). Jumba (Jumba et al., 2025) from his study also confirms that Vihiga county is an endemic region and malaria is a pressing public health challenge especially among children below five years. While other Western counties such as Kakamega also have laboratory confirmed cases records, Vihiga's high incidence and elevated malaria burden among children below five years motivated its selection for this study.

A study by Aychiluhm (Aychiluhm et al., 2020) investigated the determinants of malaria in under five children using Bayesian multilevel logistic regression model, while Mbishi (Mbishi et al., 2024) examined the prevalence of malaria among children below five years using modified poisson regression analysis. Although these studies focused on determinants and prevalence of malaria in children below

five years, neither accounted for underlying symptom progression nor incorporate state duration. Similarly, Mbete and Nyongesa (Mbete & Nyongesa, 2021) utilized Hidden Markov Model, which focuses on transition probabilities only without incorporating the time taken in any given state. The HMM implicitly assumes geometrically distributed state durations, meaning the probability of leaving a state is independent of how long the child has already been in it, and the most likely stay in any state is a single time step (Koslik, 2025). Yet the risk of a child's illness worsening may rise with the number of days already ill (Mousa et al., 2020). The Hidden Semi-Markov Model (HSMM) addresses this limitation by modelling state durations explicitly. Therefore, vital information can be obtained from Vihiga Teaching and Referral Hospital records, which will provide a better understanding of malaria symptoms progression using HSMM, enhancing early diagnosis and treatment. Hence, the focus of the study.

1.1 Statement of the Problem

Recent studies indicate that, even with many efforts to combat malaria, it is still a deadly illness and the main cause of death and morbidity in children below five years. According to WHO's 2024 World Malaria Report, malaria cases in 2023 were estimated at 263 million globally, with 597,000 deaths, of which 76% occurred in children below five years (WHO, 2024). Africa accounted for 94% of all malaria cases worldwide, estimated at 246 million in 2023, with 569,000 deaths, 78.1% of which occurred among children below five years. In Kenya, the disease caused 3.3 million cases and 11,478 deaths, with 80% of all these cases reported among children below five years (Lacey et al., 2023). Research has focused on factors contributing to malaria disease (Aychiluhm et al., 2020), malaria prevalence among children below five years (Mbishi et al., 2024), and trends in malaria morbidity and mortality. In both, each child contributes a single test result, the outcome is whether the child tested positive, and the predictors are household and environmental factors such as bed-net use, altitude and wealth. These models can identify who is at risk of infection but cannot describe symptom progression or the time spent in a state. Other studies have modeled malaria symptoms using the Hidden Markov Model (Mbete & Nyongesa, 2021) on a student population. This model estimates transition probabilities between states but assumes state durations follow a geometric (memoryless) distribution, an assumption that does not always hold in practice. However, modeling malaria symptoms in children below five years using the Hidden Semi-Markov Model (HSMM), which allows state-duration distributions to be estimated directly from data, has not been thoroughly investigated, and yet these symptoms are common and their progression rapid. As a result, this study sought to formulate a Hidden Semi-Markov Model with explicit state duration for malaria symptoms in children below five years.

1.2 Research Objective

The main aim of this study is to formulate a Hidden Semi-Markov Model, with explicit state duration, for malaria symptoms in Children below five years.

2 Literature Review

2.1 Theoretical Review

2.1.1 The Hidden Semi-Markov Model (HSMM)

This study is anchored on the Hidden Semi-Markov Model, a probabilistic model that allows the latent (underlying) process to be a Semi-Markov chain (Yu, 2010). HSMM is an extension of the Hidden Markov Model (HMM): where HMM assumes state duration is geometrically distributed and hence memoryless, HSMM removes this assumption by allowing the duration (D) in each state to be modelled using various statistical distributions estimated from the data itself, thus providing a more accurate representation of state durations. This limitation of the standard HMM has been formally established in the literature, Langrock and Zucchini (Langrock & Zucchini, 2011) demonstrated that the geometric dwell-time assumption does not always hold, and proposed allowing arbitrary state dwell-time distributions instead, a principle this study adopts. By accommodating different dwell-time distributions, HSMM can capture the true nature of symptoms more effectively, leading to better detection, characterization, and prediction of symptom patterns.

In the context of this study, HSMM can be characterized with the following parameters, which structure both the empirical review that follows and the Findings and Discussion sections of this paper:

- (i) $\pi = \{\pi_0(1), \dots, \pi_0(n)\}$ - Initial state distribution vector
- (ii) A^o - transition probability matrix between different states
- (iii) B - observation probability matrix (probabilities of the observed symptoms given the state)
- (iv) $p_j(d)$ - duration distribution of each state j , which gives the time a child spends in the state (state duration, D), with parameters μ_j and r_j for the negative binomial).

Hence, an HSMM can be defined with four parameters: $\lambda = (A^o, B, p_j(d), \pi)$

2.2 Empirical Review

Malaria disease studies in children below five years have largely applied statistical models aimed at identifying risk factors or forecasting disease trends, rather than modeling how symptoms themselves evolve over time. Aychiluhm (Aychiluhm et al., 2020) investigated the determinants of malaria in under-five children using Bayesian multilevel logistic regression, while Mbishi (Mbishi et al., 2024) examined the prevalence of malaria among children below five years using modified Poisson regression. Similarly, Gaston and Ramroop (Gaston & Ramroop, 2020), used a Generalized Additive Mixed Model (GAMM) to determine whether geographic, socio-economic and economic variables are related to malaria in children under five, while Gondwe (Gondwe et al., 2021) applied a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to forecast the trends of malaria mortality and morbidity among children below five years in Nsanje district. Although these studies differ in the statistical techniques applied, they confirm the disease burden in this age group and identify important risk factors, but none of them models disease progression over time or accounts for the underlying symptom progression that this study estimates through the hidden states (S) and observed symptoms (O).

Mbete and Nyongesa (Mbete & Nyongesa, 2021) developed a state transition model for malaria symptoms using Hidden Markov Model, estimating parameters using Bayesian method and Markov Chain Monte Carlo (MCMC) via Gibbs sampling algorithm, and confirmed a strong correlation between malaria disease and the symptoms observed. Among the studies reviewed, this is the only one identified that models malaria symptoms progression as a state transition model, and it therefore comes closest to the present study. However, it shares the HMM's limitation established in Section 2.1.1: it estimates the transition probabilities (A) between states while assuming that state durations are geometrically distributed, an assumption that does not always hold. It can therefore describe how likely a child is to move from one state to another, but not how long the child remains in that state, the parameter (D) that HSMM is designed to estimate. Several studies have applied HSMM successfully wherever the state duration aspect matters. Oliveira (Oliveira et al., 2019) developed a Hidden Semi-Markov Model to segment heart sound signals into their main components, proposing an algorithm designed to infer more suitable sojourn time parameters using a maximum likelihood approach. They tested the algorithm on three datasets: the Pascal dataset, the openly accessible PhysioNet dataset and a recently published children's dataset, using maximum likelihood criteria to determine which duration distribution best fit the data in each case. The results showed a substantial improvement in segmentation efficiency, attaining an average F-score of 92% on the PhysioNet dataset compared to 89% by previous methods. This study adopts a comparable approach to select the best fitting duration distributions across all four disease states.

Trujillo-Barreto (Trujillo-Barreto et al., 2024) modeled brain state duration in electroencephalography (EEG) data using HSMM, aiming to identify and describe changes in brain states over time. The

Hidden Semi-Markov Model (HSMM) they formulated addressed HMM's limitation by explicitly predicting the duration of brain states, rather than assuming a fixed memoryless duration. They estimated the HSMM parameters and evaluated the model's effectiveness relative to HMM on simulated data using a Bayesian approach and the variational Bayes method. Their findings indicated that HSMM is a powerful tool for EEG-driven neural state duration characterization and that it overcomes HMM's limitation, while also revealing that precise simulation of state durations is vital for sound inferences.

Huo (Huo et al., 2022) proposed a bidirectional Hidden Semi-Markov Model for electrocardiogram (ECG) waveform segmentation in cardiovascular diseases (CVDs). The study trained the model parameters using logistic regression. The study emphasized that precise delineation of ECG waves is vital for identifying subtle pathological changes in cardiac rhythm, and the findings of the experiment confirmed that bidirectional HSMM demonstrates a strong ability to segment both resting and dynamic electrocardiography (DCG) waveforms, which is helpful for identifying and monitoring CVDs. Collectively, these studies establish that HSMM consistently outperforms HMM by explicitly modelling how long a state lasts rather than assuming it, across three unrelated fields of application: heart sound, brain, and cardiac electrical signals. That the same advantage appears wherever it has been tested suggests it is a property of HSMM itself rather than a result specific to any one of these datasets. In the context of malaria, this is clinically vital: knowing how long a child is likely to remain at a given severity level supports earlier diagnosis and timely treatment, reducing the risk of rapid progression to a severe state where the risk of death is high. Despite extensive literature on malaria among children below five years, the use of Hidden Semi-Markov Model within this vulnerable group remains limited. Children in this age bracket are highly susceptible to severe malaria and contribute to approximately 76% of worldwide malaria deaths (Venkatesan, 2025). Most studies, including those by Gondwe (Gondwe et al., 2021), Gaston & Ramroop (Gaston & Ramroop, 2020), and Mbishi (Mbishi et al., 2024), focused on children below five years using statistical models such as SARIMA used for forecasting malaria trends and generalized additive model for examining socio-economic and demographic determinants of malaria. However, none models how malaria symptoms progress in children, the probabilities of moving between disease states, or the time spent in each. Similarly, Mbete and Nyongesa (Mbete & Nyongesa, 2021) utilized Hidden Markov Model, which focuses on transition and observation probabilities but assumes that the time spent in each state follows a geometric distribution. HSMM applications with explicit duration modeling have been validated in cardiac and neural signal processing (Oliveira et al., 2019; Trujillo-Barreto et al., 2024; Huo et al., 2022), but not in malaria or in pediatric symptom progression. No study reviewed here estimates all four parameters of $\lambda = (A^\circ, B, p_j(d), \pi)$ together for malaria symptom progression in children below five years. This study fills this gap through the formulation of HSMM for malaria symptoms among children below five years, in which the best-fitting duration distribution is selected for each state and used to estimate the time spent in that state. The

estimated durations show how long a child remains in a given disease state, which provides a deeper understanding of symptom progression and can inform treatment intervention. This understanding helps to determine whether a child's condition is improving or worsening. Such information can assist health practitioners to respond quickly with appropriate treatment, which may lead to lower risk of death and better recovery outcomes among children who have tested positive for malaria. Therefore, the need to reduce the mortality and morbidity rate in children within this age group due to malaria creates the importance of conducting this research.

3 Methodology

3.1 Source of Data

This study utilized secondary data on malaria symptoms among children below five years. The data were obtained from patient records at Vihiga Teaching and Referral Hospital covering the months of January through December 2023. The data were longitudinal and were restricted to this full calendar year because it captures both peak and low seasons of malaria transmission, which supports the representativeness of the data for the annual malaria burden in the region (Apat, 2022). Patients were included if they were below five years, had been tested for malaria, and had repeated observations. Those who did not meet these criteria were excluded, which gave a final sample of 300 children. Additionally, the year 2023 corresponds with the period reported in the WHO Malaria Report 2024 (WHO, 2024), the most current global malaria surveillance report available at the time of data collection. Vihiga County was selected for this study due to its high incidence and elevated malaria burden relative to other endemic counties in Western Kenya.

3.2 Observed symptoms

Zambare (Zambare et al., 2019) states that following an effective bite by an infected female *Anopheles* mosquito, symptoms occur between 7-14 days before a person can become infectious. The symptoms vary depending on individuals' immunity, age, environment and the type of malaria parasite. Martins (Martins et al., 2015) lists common symptoms of malaria including: pallor, jaundice, chills, coryza, headache, bitter mouth, myalgia, abdominal pain, sore throat, vomiting, diarrhoea, sneezing, sweating, nausea, arthralgia, dizziness, cough, and fever. Although headache, abdominal pain and jaundice are common symptoms of malaria, they were not recorded in the secondary hospital records obtained from Vihiga County Teaching and Referral Hospital and were therefore excluded from the analysis since this study relied on retrospectively collected clinical records. This, the symptoms used in this study were: y1 - body temperature, y2 - vomiting, y3 - cough, y4 - rigors, y5 - sweating, y6 - pallor, y7 - convulsion, y8 - diarrhoea and y9 - prostration.

3.3 Assumptions of the model

The following are assumptions used in the development of the model:

- (i) A child's severity is unknown (hidden) and takes one of four states (healthy, mild, moderate, severe), and the probability of moving to the next state, once the child leaves the current one, depends only on the current state.
- (ii) The time a child spends in a state follows a state-specific discrete distribution (Poisson or negative binomial), selected from the data, so the chance of leaving a state depends on how long the child has been in it.

3.4 The model

According to Yu (Yu, 2010) HSMM is an extension of Hidden Markov Models that allows the underlying state process to be a Semi Markov Chain with the aspect of state duration. State duration refers to the length of time a child remains in a specific latent state before transitioning to another. It is obtained from state duration distribution $p_j(d) = P(D_j = d)$ for $d = 1, \dots, D$, where the state duration D_j is the time taken in state j with $p_j(d) \geq 0$ and $\sum_{d=1}^D p_j(d) = 1$. Each is either a Poisson or a negative binomial distribution. Within HSMM, every state produces a sequence of observations, and the overall count of observations generated depends on the length of time taken in state i .

Assume the current time is t , we can then define the chance of transitioning from state i having stayed there for duration h , to a different state $j \neq i$ with duration d as:

$$a_{(i,h)(j,d)} = P[S_{(t+1:t+d)} = j | S_{(t-h+1:t)} = i] \quad (1)$$

This probability is taken to be invariant with respect to t for $i, j \in S$, $h, d \in D$. It shows that the predecessor state i starts at time $t - h + 1$ and ends at time t , having lasted duration h . The process then moves into state j , which starts at time $t + 1$ and runs through $t + d$. This confirms that both the entered state and its duration rely on the predecessor state and on that predecessor's own duration. Also, Equation (1) satisfies the condition $\sum_{(j \in S_i)} \sum_{(d \in D)} a_{((i,h),(j,d))} = 1 \forall i \in S$ and $h \in D$ with $a_{(i,h)(j,d)} = 0$, meaning a state that starts at time t cannot transition back into itself at time $t + 1$. This happens because state durations are governed by clearly defined distributions rather than relying on memoryless geometric distributions.

To aid our discussion on the developed symptom progression model in children below five years we used Figure 1. Before developing HSMM, we must first develop the standard HMM which is the foundation. This foundational HMM is designed to show the transitions between observed symptoms

and hidden states over time. This step is vital since HMM defines the main structure of states and the observed symptoms before HSMM can introduce the state durations. Figure 1 shows the visual representation of this process.

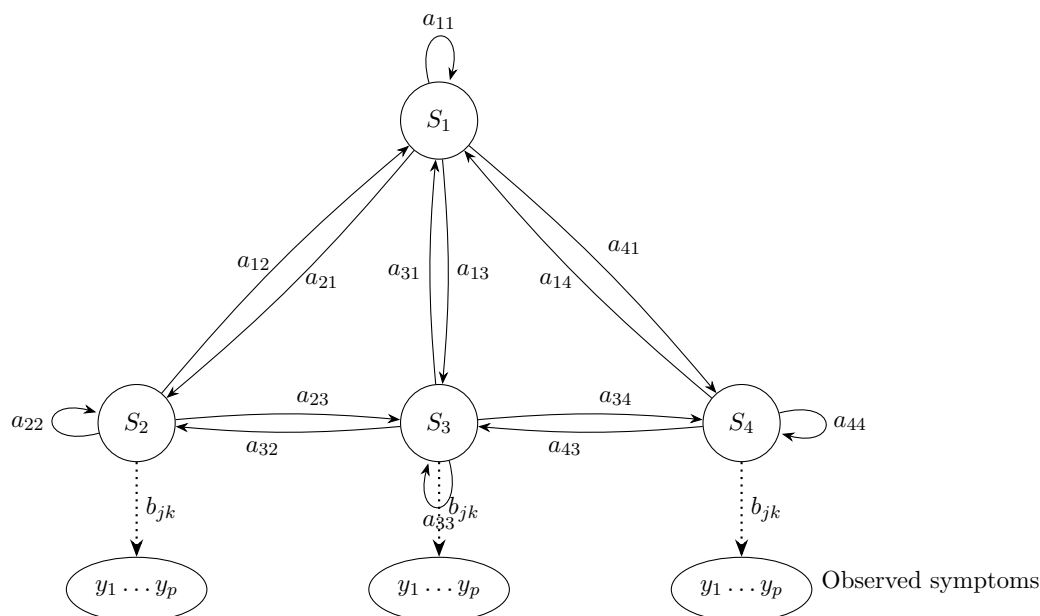


Figure 1: Hidden Markov Model

From Figure 1, we let S represent the hidden disease states, defined as $S = \{S_1, S_2, \dots, S_n\}$ be corresponding to the possible diseases states that a child may be in any given time. For purposes of this study, these latent states were categorized into four distinct states i.e: $S = \{S_1, S_2, S_3, S_4\}$. S_1 denotes the healthy/no symptoms state, S_2 the mild state, S_3 the moderate state and S_4 the severe state. Let O denote the observed symptoms defined as $O = [y_1, \dots, y_p]$ where $\{Y_i\}$'s is the different types of symptoms observed and p is total number of observed symptoms used in this study. In this study, we used nine different types of commonly observed symptoms from the health records. The symptoms used include; y_1 - body temperature, y_2 - vomiting, y_3 - cough, y_4 - rigors, y_5 - sweating, y_6 - pallor, y_7 - convulsion, y_8 - diarrhea and y_9 - prostration. Let $\pi = \{\pi_0(1), \dots, \pi_0(n)\}$ be the initial state probability vector; in which π provides the probability of beginning in each disease state. This vector is represented as an edge leading into a disease state from an origin point (state zero), which does not appear in Figure1. This is because we assume that every state has a starting point (state zero) and the system can transition out of state zero but never back into it, i.e. $\sum_s \pi_i = 1$

Let A° be the transition probability matrix, defined as a matrix whose elements describe the probabilities of moving from one state to another with the main diagonal entries set to zero. For this research, four

hidden states are used, and the matrix can be expressed as:

$$A^{\circ} = \begin{matrix} & \begin{matrix} S_1 & S_2 & S_3 & S_4 \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{matrix} & \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \end{matrix}$$

Let B represent the observation probability matrix defined as, $B = b_j(k)$. For this study, 9 distinct types of symptoms were used, thus B can be defined as:

$$B = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{matrix} & \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} & b_{1,5} & b_{1,6} & b_{1,7} & b_{1,8} & b_{1,9} \\ b_{2,1} & b_{2,2} & b_{2,3} & b_{2,4} & b_{2,5} & b_{2,6} & b_{2,7} & b_{2,8} & b_{2,9} \\ b_{3,1} & b_{3,2} & b_{3,3} & b_{3,4} & b_{3,5} & b_{3,6} & b_{3,7} & b_{3,8} & b_{3,9} \\ b_{4,1} & b_{4,2} & b_{4,3} & b_{4,4} & b_{4,5} & b_{4,6} & b_{4,7} & b_{4,8} & b_{4,9} \end{bmatrix} \end{matrix}$$

Consequently, the HSMM can now be written as: $\lambda = (A^{\circ}, B, p_j(d), \pi)$

3.4.1 Forward and backward algorithm

To derive the recursive formulas for the Forward-Backward Algorithm within the Hidden Semi Markov Model (HSMM), we define two key variables: the forward variable $\alpha_t(j, d)$, which represents the joint probability that the process remains in state j for exactly d consecutive time steps ending at time t while observing the partial sequence $O_{1:t}$, and the backward variable $\beta_t(j, d)$, which captures the chance of observing the remaining sequence $O_{t+1:T}$ given that state j has been occupied for d steps ending at time t . Formally, these are defined as: $\alpha_t(j, d) = P(S_{[t-d+1:t]} = j, O_{1:t} | \lambda)$ and $\beta_t(j, d) = P(O_{t+1:T} | S_{[t-d+1:t]} = j, \lambda)$

Based on the Markov chain property, future observations depend solely on the present state e.g

$$P(O_{t-d+1:t} | S_{[t-d-h+1:t-d]} = i, S_{[t-d+1:t]} = j, \lambda) = P(O_{t-d+1:t} | S_{[t-d+1:t]} = j, \lambda) \quad (2)$$

and

$$P(O_{t+1:T} | S_{[t-d+1:t]} = j, S_{[t+1:t+h]} = i, \lambda) = P(O_{t+1:T} | S_{[t+1:t+h]} = i, \lambda) \quad (3)$$

and are conditionally independent of earlier observations, given the present state e.g

$$P[S_{[t-d+1:t]} = j, O_{t-d+1:t} \mid S_{[t-d-h+1:t-d]} = i, O_{1:t-d}, \lambda] = P[S_{[t-d+1:t]} = j, O_{t-d+1:t} \mid S_{[t-d-h+1:t-d]} = i, \lambda] \quad (4)$$

and

$$P [O_{t+h+1:t} \mid S_{[t+1:t+h]} = i, O_{t+1:t+h}, \lambda] = P [O_{t+h+1:t} \mid S_{t+1:t+h} = i, \lambda] \quad (5)$$

We can obtain the forward and backward algorithm using Equations (2)-(5) as shown below:

$$\alpha_t(j, d) = \sum_{i \neq j; h} P [S_{[t-d-h+1:t-d]} = i, S_{[t-d+1:t]} = j, O_{1:t} \mid \lambda] \quad (6)$$

Let $A = P[S_{[t-d-h+1:t-d]} = i, O_{1:t}]$, $B = (S_{[t-d+1:t]} = j, O_{1:t})$. Applying the chain rule of probability, i.e., $P(A, B) = P(A|B) \times P(B)$ Equation (6) becomes

$$\alpha_t(j, d) = P[S_{[t-d-h+1:t-d]} = i, O_{1:t-d} \mid \lambda] \cdot P[S_{[t-d+1:t]} = j, O_{t-d+1:t} \mid S_{[t-d-h+1:t-d]} = i, O_{1:t-d}, \lambda] \quad (7)$$

But $P[S_{[t-d-h+1:t-d]} = i, O_{1:t-d} \mid \lambda] = \alpha_{t-d}(i, h)$ and from Equation (4) the $P[S_{[t-d+1:t]} = j, O_{t-d+1:t} \mid S_{[t-d-h+1:t-d]} = i, O_{1:t-d}, \lambda] = P[S_{t-d+1:t} = j, O_{t-d+1:t} \mid S_{[t-d-h+1:t-d]} = i, \lambda]$

Then Equation (7) becomes

$$\alpha_t(j, d) = \sum_{i \neq j; h} \alpha_{t-d}(i, h) P [S_{[t-d+1:t]} = j, O_{[t-d+1:t]} \mid S_{[t-d-h+1:t]} = i, \lambda] \quad (8)$$

The $P[S_{[t-d+1:t]} = j, O_{[t-d+1:t]} \mid S_{[t-d-h+1:t]} = i, \lambda]$ in Equation (8) can be simplified using the law of probability as;

$$P [S_{t-d+1:t} = j, O_{t-d+1:t} \mid S_{t-d-h+1:t} = i, \lambda] = P (S_{t-d+1:t} = j \mid S_{t-d-h+1:t-d} = i, \lambda) \cdot P (O_{t-d+1:t} \mid S_{t-d+1:t} = j, S_{t-d-h+1:t} = i, \lambda) \quad (9)$$

but $P (S_{t-d+1:t} = j \mid S_{t-d-h+1:t-d} = i, \lambda) = a_{(i,h)(j,d)}$ and the $P (O_{t-d+1:t} \mid S_{t-d+1:t} = j, S_{t-d-h+1:t} = i, \lambda) = P (O_{t-d+1:t} \mid S_{t-d+1:t} = j, \lambda)$

Now Equation (8) becomes:

$$\alpha_t(j, d) = \sum_{i \neq j; h} \alpha_{t-d}(i, h) \cdot a_{(i,h)(j,d)} P [O_{[t-d+1:t]} \mid S_{[t-d+1:t]} = j, \lambda] \quad (10)$$

but $P[O_{[t-d+1:t]} \mid S_{[t-d+1:t]} = j, \lambda] = b_{(j,d)}(O_{t-d+1:t})$

Equation (10) is now

$\alpha_t(j, d) = \sum_{i \neq j; h} \alpha_{t-d}(i, h) \cdot a_{(i,h)(j,d)} b_{(j,d)}(O_{t-d+1:t})$ (11) The same steps are followed in the backward algorithm, and we defined the backward variable as follows:

$$\begin{aligned} \beta_t(j, d) &= \sum_{i \neq j; h} P [S_{[t+1:t+h]} = i, O_{[t+1:T]} \mid S_{[t-d+1:t]} = j, \lambda] \\ &= \sum_{i \neq j; h} a_{(i,h)(j,d)} P [O_{[t+1:T]} \mid S_{[t+1:t+h]} = i, \lambda] \quad (12) \\ &= \sum_{i \neq j; h} a_{(i,h)(j,d)} \cdot b_{i,h}(O_{t+1:t+h}) \cdot P [O_{[t+1+h:T]} \mid S_{[t+1:t+h]} = i, \lambda] \end{aligned}$$

$$= \sum_{i \neq j; h} a_{(i,h)(j,d)} \cdot b_{i,h}(O_{t+1:t+h}) \cdot \beta_{t+h}(i, h)$$

In HSMM, how likely the process is to move from one state to another at time t depends on how long it has spent in that particular state before time t . Let the individual states be $S = \{S_1, S_2, \dots, S_n\}$ and the state at time t as s_t . We then define the Semi-Markov property as:

$$P(s_{t+1}=S_i \mid s_t = S_j, \dots, s_1 = S_k) = P(s_{t+1}=i \mid s_t = j, d_t(j)): 1 \leq i, j, k \leq N \quad (13)$$

Where $d_t(j)$ is defined as the length of time the process has stayed in state S_j up to time t . Let the state duration d_t , expressed as vector, be defined as:

$$d_t = \begin{cases} d_t(j) & \text{if } s_t = S_j \\ 1 & \text{if } s_t \neq S_j \end{cases} \quad (14)$$

Then the duration variable $d_t(j)$ can be obtained through induction from $d_{t-1}(j)$ as follows:

$$d_t(j) = s_t(j) \cdot s_{t-1}(j) \cdot d_{t-1}(j) + 1 \quad (15)$$

where $s_t(j)$ is 1 if $s_t = S_j$, 0 otherwise. Suppose the system is currently in state s_i at time t ; the transition matrix that depends on duration is then defined as:

$$A_{dt} = [a_{ij}(d_t)] \quad (16)$$

where $a_{ij} = P(s_{t+1} = S_j \mid s_t = S_i, d_t(i)): 1 \leq i, j \leq N$

The matrix A_{dt} can be split into two components; the recurrent state and non-recurrent state transition probabilities. Let $P_{ii}(d)$ be how likely the process is to remain in state i for exactly d time units. If the duration is geometric, as implied by an HMM {cite} $P_{ii}(d)$ is obtained as: $P_{ii}(d) = a_{ii}^{d-1} (1 - a_{ii})$, $E[d] = \frac{1}{1-a_{ii}}$ (17)

where $P_{ii}(d)$ gives the probability of the process staying in the same state for d consecutive time instances.

In HSMM, geometric duration (Equation 17) is replaced by the distribution selected for each state, so that $P_{ii}(d)$ is replaced by $p_j(d)$, the probability that a child stays exactly d days in state j , for $d = 1, 2, \dots, D$, rescaled so that $\sum_{d=1}^D p_j(d) = 1$. Since durations start at one day, the Poisson or Negative Binomial distributions, whichever is selected by AIC, is shifted by one day. For the shifted Poisson distribution,

$$p_j(d) = \frac{e^{-\mu_j} \mu_j^{d-1}}{(d-1)!}, \quad E[d_j] = \mu_j + 1 \quad (18)$$

and for the shifted Negative Binomial distribution,

$$p_j(d) = \frac{\Gamma(d-1+r_j)}{(d-1)!\Gamma(r_j)} q_j^{r_j} (1-q_j)^{d-1}, \quad q_j = \frac{r_j}{r_j+\mu_j}, \quad E[d_j] = \mu_j + 1 \quad (3.43) \quad (19)$$

where μ_j is the mean of the shifted duration $d - 1$ of state j , r_j is the dispersion parameter, and $E[d_j]$ is exact when D is large enough for the truncated tail to be negligible. Unlike the geometric case, the probability of leaving state j after d days is not constant for the Poisson distribution, nor the Negative Binomial distribution unless $r_j = 1$, so the HSMM does not reduce to an HMM (Yu, 2010).

On the other hand, non-recurrent transition probability is the transition occurring between two distinct states. It is defined as $A^\circ = [a_{ij}^\circ]$ and represented by a matrix in which the diagonal entries are set to zero, given as:

$$A^\circ = [a_{ij}^\circ] = \begin{cases} 0 & \text{if } i = j \\ P(s_{t+1} = S_j \mid s_t = S_i) & \text{if } s_t \neq S_j \end{cases} \quad (20)$$

and A° must satisfy the condition $\sum_{j=1}^N a_{ij}^\circ = 1 \forall i$.

3.4.2 State duration distribution

State duration is how long a child stays in a disease state before progressing to another. One of the assumptions of HMM is that the time taken in a specific state follows a geometric distribution, however sometimes this assumption does not hold (Langrock & Zucchini, 2011). HSMM has become a powerful tool that overcomes this limitation by allowing the underlying state to be a semi-markov chain which takes arbitrary distributions except geometric and exponential distributions.

The expected duration $E[d] = \frac{1}{1-a_{ii}}$, in Equation 17 follows a geometric distribution describing a standard HMM behaviour (Rabiner, 1989). The quantity is obtained directly from the transition matrix rather than estimated from the data. To obtain the explicit state duration for the four states, state-specific run-lengths were obtained from the Viterbi-decoded sequence: that is the number of consecutive clinical observations classified into the same state (Koslik, 2025). This gives empirical state durations obtained from the data rather than assumed. Dispersion index (variance-to-mean-ratio (VMR)) was calculated for each state (National Institute of Standards and Technology, 2017). This is because HMM enforces geometric distribution. A value close to 1 indicates equidispersion, consistent with the geometric distribution's memoryless assumption. On the other hand, a value greater than 1 indicates overdispersion, and a value less than 1 indicates underdispersion (National Institute of Standards and Technology, 2017). The $VMR > 1$ and $VMR < 1$ shows that, HMM is inadequate hence the need for more flexible distributions.

Poisson and Negative Binomial distributions were fitted to each state's run-length data using maximum likelihood because both are natively discrete (Guédon, 2003), avoiding arbitrary parameters such as the fixed trial count (n) required by a binomial distribution (Lloyd-Smith, 2007). Other candidate distributions were explicitly excluded: the Hypergeometric distribution was rejected because it models finite sampling without replacement, the Bernoulli distribution was excluded because it cannot represent

durations exceeding a single day (Feller, 1991); and the Discrete Uniform distribution was ruled out because it assumes transition risks have no time dependence, which lacks a clinical basis (Yu,2010). Model selection between the two distributions was carried out using AIC, $AIC = -2 \ln L + 2k$ (Akaike, 1974), where $\ln L$ is the maximized log-likelihood and k is the number of free parameters. A lower AIC value indicates the better fitting distribution for that state, and this is compared to the dispersion index to determine whether an explicit-state duration model is required. This approach allows state duration for malaria severity states to be determined by the data itself.

4 Findings & Discussion

4.1 Distribution of Malaria Severity States (S)

Table 1 shows the distribution of malaria severity over all observations and for the total number of children.

Table 1

Distribution of Malaria severity

Malaria Severity	No. of observations	% of observations	No. of children	% of children
Healthy	305	23.81	63	21.000
Mild symptoms	273	21.31	42	14.00
Moderate symptoms	286	22.33	49	16.33
Severe symptoms	417	32.55	146	48.67
Total	1281	100.00	300	10.00

Table 1, shows the distribution of malaria severity. Of the 1,281 observations, 23.81% were healthy, 21.31% mild, 22.33% moderate and 32.55% severe. Table 1 also shows that 21% of the children examined at the hospital after displaying malaria related symptoms tested negative for malaria. Results further indicate that 42 children (14%) presented with mild symptoms, 49 (16.3%) with moderate symptoms and 146 (48.7%) with severe symptoms. This tells us that severe was the most common, followed by no malaria, moderate and lastly mild.

4.1.1 Profile of Observed Symptoms (O)

Table 2 shows the nine observed symptoms with their corresponding means, standard deviations and standard errors.

Table 2

Licensed Under Creative Commons Attribution (CC BY-NC)

Descriptive Statistics

Variable	%	Mean	Std. Dev.	S.E.
Fever	94.3	1.792642	0.914359	0.052879
Vomiting	85.3	1.892977	1.050048	0.060726
Cough	70.6	1.511706	1.213333	0.070169
Rigors	78.9	1.692308	1.200800	0.069444
Sweating	77.9	0.782609	0.421206	0.024359
Pallor	81.6	1.779264	1.089042	0.062981
Convulsion	75.6	1.722408	1.258474	0.072779
Diarrhoea	66.9	1.227425	1.084570	0.062722
Prostration	77.6	1.882943	1.216364	0.070344

Results from Table 2 show that fever, vomiting, and pallor were the most frequently observed symptoms, at 94.3%, 85.3%, and 81.6% respectively, while diarrhoea was the least observed at 66.9%. The means across symptoms are fairly close to one another; fever, for instance, had an average value of 1.792642 and a standard deviation of 0.914359.

The low standard error (SE) values obtained indicate that the sample means accurately represents the true population means. Pallor, for example, recorded a mean of 1.779264 and a standard error of 0.062981, whereas rigors recorded a mean of 1.692308 and a standard error of 0.069444. The standard deviation indicates the extent to which individual values deviate from the mean. For Diarrhoea, the mean score was 1.23, with a standard deviation of 1.0845, showing that individual scores deviated from the mean by approximately 1.0845 points on average.

4.2 Estimating the Initial State Distribution (π)

The initial state distribution (π) denotes the probability of a child beginning in each disease state. The estimated values are presented below:

$$\pi = \begin{bmatrix} 0.2049 \\ 0.1452 \\ 0.1594 \\ 0.4905 \end{bmatrix}$$

where $\sum \pi = 1.0000$

These estimated values are 0.2049 for S_1 (no malaria), 0.1452 for S_2 (mild), 0.1594 for S_3 (moderate)

and 0.4905 for S_4 (severe), which sum to 1. S_4 is therefore the most likely starting state: about 49% of children are estimated to begin in severe malaria, compared with about 20% in S_1 , 15% in S_2 and 16% in S_3 . These values describe only where a sequence starts.

4.3 Estimating the State Transition Probabilities (A°)

To compute the transition and emission probabilities from the observed symptom sequences, the model takes the emission sequence (i.e. observed symptoms) along with the hidden states, then yields estimates for both the transition and observation probabilities. The resulting optimal value determined from the transition matrix, A° , between malaria disease states is presented below;

$$A^\circ = \begin{matrix} & \begin{matrix} S_1 & S_2 & S_3 & S_4 \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{matrix} & \begin{bmatrix} 0.0000 & 0.2188 & 0.2188 & 0.5624 \\ 0.3077 & 0.0000 & 0.1153 & 0.5770 \\ 0.4643 & 0.0357 & 0.0000 & 0.50000 \\ 0.1000 & 0.4500 & 0.4500 & 0.0000 \end{bmatrix} \end{matrix}$$

The transition matrix A° gives the probability of moving from one state to another. The diagonal entries are zero ($a_{11} = a_{22} = a_{33} = a_{44} = 0$) because in an HSMM the time spent in a state is modelled explicitly by a duration distribution, so self-transitions are excluded and the diagonal of the transition matrix is zero (Yu, 2010).

For children in the healthy state, there is a 22% chance ($a_{12} = 0.2188$) of moving to the mild state, a 22% chance ($a_{13} = 0.2188$) of transitioning to the moderate state, and a 56% chance ($a_{14} = 0.5624$) of moving directly to the severe state. This suggests that children leaving the healthy state are more likely to move directly to severe malaria than to the mild or moderate states. Similarly, for children in the mild state, there is a 31% chance ($a_{21} = 0.3077$) of recovering and moving to the healthy state, a 12% chance ($a_{23} = 0.1153$) of transitioning to the moderate state, and a 58% chance ($a_{24} = 0.5770$) to the severe state. This indicates that children leaving the mild state are more likely to deteriorate to the severe state than to recover or move to the moderate state. In addition, children in the moderate state have a 46% chance ($a_{31} = 0.4643$) of recovering and transitioning to the healthy state, a 4% chance ($a_{32} = 0.0357$) of moving to the mild state, and a 50% chance ($a_{34} = 0.5000$) moving to the severe state. This implies that children in the moderate state have nearly equal chances of recovering and progressing to severe malaria, although progression to the severe state is slightly more likely. While, for children in the severe state, there is an 10% chance ($a_{41} = 0.1000$) of recovering, a 45% chance ($a_{42} = 0.4500$) of progressing to the mild state, and a 45% chance ($a_{43} = 0.4500$) of moving

to the moderate state. This means that once a child reaches the severe state, they mostly step down to the mild or moderate state, while a direct return to the healthy state is less common. Generally, the findings from the transition matrix A° tells us how likely a child is to transition from one state to another.

4.4 Estimating the Observation (Emission) Probabilities (B)

The nine symptoms considered in this study were: y_1 - body temperature, y_2 - vomiting, y_3 - cough, y_4 - rigors, y_5 - sweating, y_6 - pallor, y_7 - convulsion, y_8 - diarrhea and y_9 - prostration.

Observation probability matrix B

$$B = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{matrix} & \begin{bmatrix} 0.023 & 0.523 & 0.364 & 0.023 & 0.795 & 0.205 & 0.023 & 0.727 & 0.023 \\ 0.028 & 0.944 & 0.500 & 0.028 & 0.778 & 0.972 & 0.028 & 0.667 & 0.944 \\ 0.027 & 0.919 & 0.946 & 0.027 & 0.730 & 0.919 & 0.027 & 0.730 & 0.919 \\ 0.010 & 0.910 & 0.790 & 0.010 & 0.780 & 0.990 & 0.010 & 0.670 & 0.990 \end{bmatrix} \end{matrix}$$

Observation matrix B presents the observation probabilities, capturing the relationship between each observed symptom and the underlying hidden state. Each emission probability in matrix B represents the chance of observing a specific symptom given that a child is in a particular state. For example, pallor has a 99% probability of being observed in the severe state, whereas rigors show only a 1% probability in the same state. This shows how symptom patterns differ across disease states.

4.5 Computing the Most Likely Sequence (Viterbi-Decoded State Sequence)

The Viterbi algorithm was applied to the test set in order to identify the most probable sequence of hidden states. The decoding was conducted using the transition and emission probabilities estimated from the training data set. The findings indicate that the most likely state was S4. In order to evaluate how accurate this decoding was, the proportion of time points where the Viterbi-decoded sequence matched the actual state sequence was computed. The result showed a 94.44% agreement, indicating that the most probable sequence of states, as determined by the HSMM, closely matched the actual sequence of disease progression states in most cases.

4.6 Estimating Explicit State Durations (D)

State duration is length of time a child remains in a given disease state before transitioning to a different state.

4.6.1 Selecting the Best-Fitting State Duration Distribution

Poisson and Negative Binomial distributions were fitted to the run-lengths of the Viterbi decoded state sequence for each state, and the best fitting distribution was selected using Akaike Information Criteria (AIC).

Table 3

State Duration Distribution Selection (HSMM)

State	Dispersion Index	Poisson AIC	NegBin AIC	Best-Fit Distribution
S1 – Healthy/No Malaria	0.3523	81.6198	89.2561	Poisson
S2 – Mild	0.1128	58.9591	62.8045	Poisson
S3 – Moderate	0.3151	67.3124	73.1883	Poisson
S4 – Severe	1.6397	174.6805	151.0635	Negative Binomial

Results from Table 3 show that the healthy, mild, and moderate states each recorded a dispersion index below 1 (0.3523, 0.1128, and 0.3151, respectively), together with lower Poisson AIC values of (81.6198, 58.9591, and 67.3124) than their Negative Binomial counterparts (89.2561, 62.8045, and 73.1883). This confirms the Poisson distribution as the best fit for these three states. The severe state, by contrast, recorded a dispersion index of 1.6397, together with a lower Negative Binomial AIC of 151.0635 than Poisson AIC of 174.6805, confirming that the Negative Binomial distribution is the best fit distribution for this state. With 48 runs recorded against 32, 28, and 26 for the no-malaria, moderate, and mild states respectively, the severe state was also the most frequently occupied among the four states.

4.6.2 Estimated State Durations

Having established the best-fitting distribution for each state in Section 4.6.1, the corresponding mean state durations were computed and are presented in Table 4.

Table 4

Estimated State Duration Distribution

State	Best-Fit Distribution	Mean Duration (days)
S1 - Healthy/No Malaria	Poisson	1.956
S2 - Mild	Poisson	1.758
S3 - Moderate	Poisson	1.800
S4 - Severe	Negative Binomial	2.772

Results from Table 4 show that healthy, mild, moderate state recorded mean durations of 1.956, 1.758, and 1.800 days respectively, indicating that a child presenting in any of these states is expected to remain there for approximately two days before improving or progressing to a more severe state. On the other hand, the severe state recorded a longer mean duration of 2.772 days, showing that a child with severe symptoms will remain in that state for approximately three days.

4.7 Discussion

The finding that 48.67% of children examined were in the severe state, against only a 10% probability of transitioning from the severe state back to the healthy state, builds on the state transition model developed by Mbete and Nyongesa (Mbete & Nyongesa, 2021), who used a Hidden Markov Model to confirm a strong correlation between malaria disease and the symptoms observed. The present findings extend that correlation by quantifying the transition probabilities in both directions: the probability of transitioning from the moderate to the severe state (50%) is far higher than the probability of transitioning from the severe state back to the healthy state (10%). This directional pattern in the transition probabilities is consistent with the WHO's 2024 World Malaria Report, which attributes 76% of global malaria deaths to children below five years (WHO, 2024); the low probability of transitioning out of the severe state offers a possible explanation, why this age group accounts for such a high share of malaria deaths

The dominance of fever (94.3%), vomiting (85.3%), and pallor (81.6%) in this study's sample of children below five years is consistent with Martins et al. (2015), who identified these among the most common malaria symptoms in semi-immune Amazonian adults. The two populations differ in age and immune exposure: Martins et al.'s were adults with immunity from repeated prior infection, while under-fives are classified by WHO as having a weak immune system and the highest risk of severe outcomes (WHO, 2024). That the same three symptoms dominate in both suggests they may be consistent early indicators of malaria rather than specific to one population, though confirming this would need a direct multi-site comparison which is beyond this study's scope.

The observation matrix's finding that pallor carries a 99% probability of being observed in the severe state, against only 1% for rigors in the same state, is a symptom-to-state association that goes beyond what the determinants literature (Aychiluhm et al., 2020) or the prevalence literature (Mbishi et al., 2024) reports, since neither models how individual symptoms relate to specific states of disease progression. Zambare (Zambare et al., 2019) notes that malaria symptoms vary according to a person's immunity, age, environment, and the infecting parasite type; the present findings add a state-specific dimension to this variation, showing that among children below five years, pallor is concentrated overwhelmingly in the severe state rather than distributed evenly across states. This distinction is not visible in studies that treat symptom occurrence as a single undifferentiated measure, and it speaks

directly to this study's aim of formulating a model that captures the explicit progression of malaria symptoms. The dispersion index of 1.6397 obtained for the severe state provides direct empirical support for Langrock and Zucchini (Langrock & Zucchini, 2011), who demonstrated theoretically that the geometric dwell-time assumption underlying the standard HMM does not always hold. This finding directly contradicts what the Hidden Markov Model used by Mbete and Nyongesa (Mbete & Nyongesa, 2021) would assume for the same data: a fixed, memoryless duration in every state. Instead, the healthy, mild, and moderate states were each best described by the Poisson distribution, with dispersion index values below 1 indicating tightly clustered durations, while only the severe state required the added flexibility of the Negative Binomial distribution. This shows that the geometric assumption fails specifically where it matters most clinically, in the state associated with the highest mortality risk, a distinction the Hidden Markov Model applied by Mbete and Nyongesa (Mbete & Nyongesa, 2021) would have been unable to detect. The improvement documented here is consistent with Oliveira (Oliveira et al., 2019), who reported a rise in segmentation accuracy from 89% to 92% when duration was estimated directly from data rather than assumed. Trujillo-Barreto (Trujillo-Barreto et al., 2024) and Huo (Huo et al., 2022) similarly found that Hidden Semi-Markov Models outperformed Hidden Markov Models in EEG and ECG state-duration modelling respectively. The present findings extend this pattern, established previously only in signal processing, into pediatric infectious disease, indicating that the benefit of explicit duration modelling generalizes beyond the domains in which it was first demonstrated.

Aychiluhm (Aychiluhm et al., 2020), Mbishi (Mbishi et al., 2024), Gaston and Ramroop (Gaston & Ramroop, 2020), and Gondwe (Gondwe et al., 2021) each modeled malaria in children below five years using regression- or time-series-based methods that identify risk factors or forecast trends, but none captures how a child's condition evolves once symptoms appear. The present findings do not contradict this body of work; they answer a different question that these studies leave open. Where the earlier studies establish who is at risk, the transition and duration estimates reported here establish how quickly a symptomatic child is likely to deteriorate, translating risk-factor knowledge into a basis for setting response-time targets for community health workers, caregivers, and healthcare practitioners.

The short mean durations of approximately two days in the healthy, mild, and moderate states, against roughly three days in the severe state, quantify the rapid symptom progression associated with the high childhood mortality reported by WHO (WHO, 2024) and Lacey (Lacey et al., 2023). This finding directly highlights the need for early intervention to prevent rapid progression to severe state of malaria among children below five years where the risk of mortality is high.

5 Conclusion & Recommendations

5.1 Conclusion

An HSMM with four states (healthy, mild, moderate, and severe) was formulated and fitted to 300 children based on observed symptoms. The findings revealed that fever, vomiting, and pallor were the most frequently observed symptoms. The estimated transition probabilities did not show a step-by-step progression from healthy to mild to moderate to severe. Children leaving the healthy state moved most often directly to the severe state (0.56), and the same was true for children leaving the mild state (0.58). From the severe state, children mostly moved to the mild or moderate states (0.45 each), and a direct return to the healthy state was less common (0.10). Poisson and Negative Binomial distributions were fitted, with the better-fitting distribution selected using the Akaike Information Criterion (AIC). This analysis found that the healthy, mild, and moderate states were each well described by the Poisson distribution, with dispersion index values well below 1 indicating that durations in these states were tightly clustered around their means. In contrast, the severe state required the additional flexibility of the Negative Binomial distribution, with a dispersion index of 1.6397 indicating that individual durations varied considerably.

The estimated mean durations were 1.96 days (healthy), 1.76 days (mild), 1.80 days (moderate), and 2.77 days (severe). The shorter durations in the earlier states suggest rapid progression to the severe state, where the risk of mortality is high. Collectively, these results shift the analytical focus in pediatric malaria research away from the determinants, prevalence, and incidence of malaria and toward understanding how a child's symptoms progress over time once the disease is present, providing a data-driven basis for timely intervention. These findings demonstrate that HSMMs, by explicitly modeling state durations, effectively capture malaria symptoms progression and highlight the need for early diagnosis and targeted intervention strategies in pediatric populations.

5.2 Recommendations

Based on the findings of this study, several avenues for further research are recommended. Although this study focused on children under five years, we recommend applying the model to predict and control malaria across all age groups. The model should incorporate demographic indicators, including marital status, age, income, gender, occupation, and education level, to provide a deeper understanding of the populations at higher risk and to inform relevant control measures. In addition, the study recommends a larger multi-hospital study to assess whether the transition probabilities, emission probabilities, and state durations found here are consistent in other settings, since the model was fitted on data from a single hospital. These steps can improve understanding of malaria and support the development of strategies to reduce mortality among children under five years.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Apat, D. O. (2022). Spatio-Temporal Distribution of Fever-Related Malaria and Non-Malarial Cases among Children Aged Below Five Years and their Relationship with Rainfall, in Siaya County, Kenya (Doctoral dissertation).
- Aychiluhm, S. B., Gelaye, K. A., Angaw, D. A., Dagne, G. A., Tadesse, A. W., Abera, A., & Dillu, D. (2020). Determinants of malaria among under-five children in Ethiopia: Bayesian multilevel analysis. *BMC Public Health*, 20(1), 1468. <https://doi.org/10.1186/s12889-020-09560-1>
- Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 2, 2nd ed.). John Wiley & Sons.
- Gaston, R. T., & Ramroop, S. (2020). Prevalence of and factors associated with malaria in children under five years of age in Malawi, using malaria indicator survey data. *Heliyon*, 6(5), e03946. <https://doi.org/10.1016/j.heliyon.2020.e03946>
- Gondwe, T., Yang, Y., Yosefe, S., Kasanga, M., Mulula, G., Luwemba, M. P., Jere, A., Daka, V., & Mudenda, T. (2021). Epidemiological trends of malaria in five years and under children of Nsanje district in Malawi, 2015–2019. *International Journal of Environmental Research and Public Health*, 18(23), 12784. <https://doi.org/10.3390/ijerph182312784>
- Guédon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12(3), 604–639. <https://doi.org/10.1198/1061860032030>
- Huo, R., Zhang, L., Liu, F., Wang, Y., Liang, Y., & Wei, S. (2022). ECG segmentation algorithm based on bidirectional hidden semi-Markov model. *Computers in Biology and Medicine*, 150, 106081. <https://doi.org/10.1016/j.combiomed.2022.106081>
- Jumba, B. N., Webale, M., Makwali, J., & Shaviya, N. (2025). Predictors of complicated pediatric malaria among children under five in the Vihiga Highlands, Western Kenya. *African Journal of Empirical Research*, 6(1), 895-905. <https://doi.org/10.51867/ajernet.6.1.76>
- Koslik, J. O. (2025). Hidden semi-Markov models with inhomogeneous state dwell-time distributions. *Computational Statistics & Data Analysis*, 209, 108171. <https://doi.org/10.1016/j.csda.2025.108171>
- Lacey, H., Jain, N., Sugimoto, M., Shimato, M., Reine, I., & Oria, K. (2023). Combating malaria in Kenya through collaborative population health education: a systematic review and pilot case study. *Infectious Diseases*, 55(10), 664-683. <https://doi.org/10.1080/23744235.2023.2231082>

- Langrock, R., & Zucchini, W. (2011). Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics & Data Analysis*, 55(1), 715-724. <https://doi.org/10.1016/j.csda.2010.06.015>
- Li, T., & Ma, J. (2022). On theoretical justification of the forward-backward algorithm for the variational learning of Bayesian hidden Markov models. *IET Signal Processing*, 16(6), 674-679. <https://doi.org/10.1049/sil2.12129>
- Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, 2(2), e180. <https://doi.org/10.1371/journal.pone.0000180>
- Martins, A. C., Araújo, F. M., Braga, C. B., Guimarães, M. G. S., Nogueira, R., Arruda, R. A., Fernandes, L. N., Correa, L. R., Malafronte, R. dos S., Cruz, O. G., Codeço, C. T., & da Silva-Nunes, M. (2015). Clustering symptoms of non-severe malaria in semi-immune Amazonian patients. *PeerJ*, 3, e1325. <https://doi.org/10.7717/peerj.1325>
- Mousa, A., Al-Taiar, A., Anstey, N. M., Badaut, C., Barber, B. E., Bassat, Q., ... & Okell, L. C. (2020). The impact of delayed treatment of uncomplicated *P. falciparum* malaria on progression to severe malaria: a systematic review and a pooled multicentre individual-patient meta-analysis. *PLoS medicine*, 17(10), e1003359. <https://doi.org/10.1371/journal.pmed.1003359>
- Mbete, D. A., & Nyongesa, K. (2021). State-transition model for malaria symptoms. *Asian Journal of Probability and Statistics*, 10(4), 22–46. <https://doi.org/10.9734/ajpas/2020/v10i430253>
- Mbishi, J. V., Chombo, S., Luoga, P., Omary, H. J., Paulo, H. A., Andrew, J., & Addo, I. Y. (2024). Malaria in under-five children: prevalence and multi-factor analysis of high-risk African countries. *BMC Public Health*, 24(1), 1687. <https://doi.org/10.1186/s12889-024-19206-1>
- National Institute of Standards and Technology. (2017). Index of dispersion. NIST/SEMATECH e-Handbook of Statistical Methods. <https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/inddisp.htm>
- Odhiambo, S. O. (2023). *Climate variability and malaria transmission trends in different altitudes of Lower Lake Victoria Basin, Kenya* (Doctoral dissertation, Kenyatta University). Kenyatta University Institutional Repository. <https://ir-library.ku.ac.ke/handle/123456789/27380>
- Oliveira, J., Renna, F., Mantadelis, T., & Coimbra, M. (2019). Adaptive sojourn time HSMM for heart sound segmentation. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 642-649. <https://doi.org/10.1109/JBHI.2018.2841197>
- Ouma, F. (2021). *Malaria data quality assessment in Kakamega and Vihiga Counties*. Kenya Field Epidemiology Training Program (K-FELTP) & Division of National Malaria Program (DNMP). <http://ajernet.net>

[//www.feltp.or.ke/Malaria-Data-Quality-Assessment-In-Kakamega-And-Vihiga-Counties-September.html](http://www.feltp.or.ke/Malaria-Data-Quality-Assessment-In-Kakamega-And-Vihiga-Counties-September.html)

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286. <https://doi.org/10.1109/5.18626>

Trujillo-Barreto, N. J., Araya Galvez, D., Astudillo, A., & El-Deredy, W. (2024). Explicit modeling of brain state duration using hidden semi Markov models in EEG data. *IEEE Access*, 12, 12335–12355. <https://doi.org/10.1109/ACCESS.2024.3354711>

World Health Organization. (2024). *World malaria report 2024*. <https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2024>

Yu, S. Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2), 215–243. <https://doi.org/10.1016/j.artint.2009.11.011>

Zambare, K. K., Thalkari, A. B., & Tour, N. S. (2019). A review on pathophysiology of malaria: A overview of etiology, life cycle of malarial parasite, clinical signs, diagnosis and complications [sic]. *Asian Journal of Research in Pharmaceutical Science*, 9(3), 226–230. <https://doi.org/10.5958/2231-5659.2019.00035.3>

© Ogatenga et al. (2026).; This is an Open Access article distributed under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.