Received: 11 March 2021 | Revised: 17 May 2022 | Accepted: 10 July 2022

DOI: 10.1002/sim.9534

RESEARCH ARTICLE

Statistics in Medicine WILEY

Classification of disease recurrence using transition likelihoods with expectation-maximization algorithm

Huijun Jiang[1] | **Quefeng Li**[1] | **Jessica T. Lin**[2] | **Feng-Chang Lin**[1]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA

[2]Division of Infectious Disease, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

Correspondence
Feng-Chang Lin, Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.
Email: flin@bios.unc.edu

Funding information
National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR002489; National Institute of Allergy and Infectious Diseases, Grant/Award Number: K08AI110651; National Institute on Aging, Grant/Award Number: R01AG073259

When an infectious disease recurs, it may be due to treatment failure or a new infection. Being able to distinguish and classify these two different outcomes is critical in effective disease control. A multi-state model based on Markov processes is a typical approach to estimating the transition probability between the disease states. However, it can perform poorly when the disease state is unknown. This article aims to demonstrate that the transition likelihoods of baseline covariates can distinguish one cause from another with high accuracy in infectious diseases such as malaria. A more general model for disease progression can be constructed to allow for additional disease outcomes. We start from a multinomial logit model to estimate the disease transition probabilities and then utilize the baseline covariate's transition information to provide a more accurate classification result. We apply the expectation-maximization (EM) algorithm to estimate unknown parameters, including the marginal probabilities of disease outcomes. A simulation study comparing our classifier to the existing two-stage method shows that our classifier has better accuracy, especially when the sample size is small. The proposed method is applied to determining relapse vs reinfection outcomes in two *Plasmodium vivax* treatment studies from Cambodia that used different genotyping approaches to demonstrate its practical use.

KEYWORDS
classification, EM algorithm, infectious diseases, malaria, transition likelihood

1 | INTRODUCTION

Infectious diseases are a leading cause of death worldwide, particularly in low-income countries.[1] Recurrence of infection after treatment makes disease control difficult and is a critical factor for treatment efficacy. Two processes can cause recurrence: relapse of the primary infection or reinfection. Identifying recurrence from more than one potential cause is important in disease control. Modeling disease progression is usually the first step.

The disease progression modeling, known as DPM, utilizes mathematical functions to describe the disease's time course and tracks disease severity over time. It enables a better understanding of disease prognosis and provides insights into effective treatment. Statistical modeling of DPM and its estimation methods have been proposed and studied in the past decades. Early work in DPM, such as the Emax model[2] and path model,[3] is commonly used for studying the time course of drug effects. Desper et al[4,5] developed oncogenetic tree models and established distance-based tree models. Extending the oncogenetic tree model, Beerenwinkel et al[6] proposed a mixture model that described the complexity of disease progression via multiple trees. Gerstung et al[7] otherwise used a hidden conjunctive Bayesian network modeling

that divides the disease process into a hidden part of the actual event accumulation and an observed part of the erroneous process. Tofigh[8] also suggested an entire course of the disease as a hidden process influenced by observational errors. Multi-state Markov models in continuous time are also often used to model the course of disease progression. Bureau et al[9] proposed a hidden Markov chain approach that considered the actual disease outcome as the hidden state of a continuous-time Markov chain with an imperfect measurement as an observation. Hjelm et al[10] further generalized the tree models by suggesting probabilistic network aberration models based on Markov models.

Critical therapeutic decisions are based on an understanding of the disease state, especially when the disease recurs. For example, malaria parasites have a complex life cycle. *Plasmodium vivax* and *Plasmodium ovale*, in particular, have a latent liver stage that magnifies their epidemiological and clinical complexity. Relapses from dormant liver-stage *P vivax* parasites (hypnozoites) are responsible for the bulk of the disease burden due to vivax malaria.[11] Developing classification criteria is important in developing and testing anti-relapse drugs and designing public health interventions. In clinical trials of primaquine and other anti-relapse drugs, the absolute efficacy of a drug against relapse can only be estimated against the number of infections in the comparator group, as a certain proportion of those treated were reinfected over time (reinfection), even with 100% efficacy of the anti-relapse drug.[12,13] From a public health standpoint, if we could retrospectively determine how many patients present with relapses as opposed to reinfection, we can begin to understand the disease burden attributable to relapse and invest in campaigns to detect and treat those harboring latent liver infection.

However, it is clinically impossible to distinguish the cause of recurrent blood-stage infection from hypnozoite-derived (relapse), a blood-stage treatment failure (recrudescence), or a newly acquired infection (reinfection),[14] even though each of these requires a different prevention strategy. Until now, few methods can determine which patients who administered anti-relapse drugs failed therapy in clinical trials due to relapse vs reinfection. To classify the late treatment failures of *P vivax* as recrudescence or reinfection in Ethiopia, Plucinski et al[15] developed a Bayesian algorithm to estimate the posterior probability of a recrudescent infection using microsatellite genotyping data. Jones et al[16] later showed that the Bayesian algorithm has high accurate estimates of the true recrudescence across different transmission and drug failure rates, especially in scenarios with a high number of recrudescent patients. Taylor et al[14] used time-to-event information to derive prior probabilities for each of the three recurrent states and subsequently derived the posterior information based on a genetic model incorporating *P vivax* microsatellite marker data. Genotyping the microsatellites can also determine the recrudescence in *Plasmodium falciparum* by overlapping a variant in both initial and recurrent infections.[17] However, sharing a prevalent variant likely has a false positive relapse because a patient can also contact a new infection with the variant in the environment.[18]

In recent technology, the targeted amplicon deep sequencing has been used to differentiate reinfection from relapse.[19] It was hypothesized that through genotyping of the initial and recurrent parasite isolates, one might distinguish relapse from reinfection based on the variant overlap between two sequencing results within an individual. A transition model can naturally describe the presence or absence of variants between those two sequencing results. However, an unknown mixture of two causes, relapse and reinfection, complicates the estimation of transition probabilities. Lin et al[20] proposed a novel two-stage method to estimate the transition probabilities. They first established a statistical model to describe the relapse probability in the recurrent infection and then used baseline information (initial sequencing) to obtain the first stage estimates. The estimates are further plugged into the likelihood functions of the transition model, that is, transitional likelihoods, to get the second stage estimates. Using a ratio of two transition likelihoods, one can update the classification probability from the first stage to the second one utilizing the transition information. While this approach performs well using overlapped genetic information between disease occurrence and recurrence, it relies on large sample size for consistent estimation in both stages.

The EM algorithm[21] has been shown to have broad application in a wide variety of incomplete data problems. In this classification problem with a missing cause of the disease in every recurrence, the EM algorithm shall be more advantageous by simultaneously optimizing the joint likelihood function in two stages. This article aims to classify the unknown cause of the disease recurrence via transition information of covariates observed in both disease occurrence and recurrence. We treat the first observed disease occurrence as the baseline, regardless of whether it reflects a recurrence from a prior disease. The method is first established on a multinomial logit model that describes the likelihood of latent disease status when the disease occurs. We then update the disease outcome probability by a ratio of transition likelihoods of the baseline covariates under a mixture distribution, using a similar approach to Lin et al.[20] In addition, we relax the assumption that the transition probabilities of baseline covariates are all equal, assuming the transition probability depends on a subject-level covariate. The complete data likelihood function that incorporates the missing cause of the disease is constructed. A surrogate function that takes expectation of the complete-data likelihood function is derived in the

E-step, given observed data and the current value of parameter estimates. Maximization of the surrogate function updates the parameter estimates, and an iteration between E- and M-steps continues until the parameter estimates converge.

The rest of this article is organized as follows. In Section 2, we develop a multinomial logit model for the probability of observing a disease recurrence in the follow-up period, which sums over probabilities of all latent outcomes of the disease. An EM algorithm is developed to estimate the parameter of interest. A more accurate classifier using both baseline and recurrence covariates information is derived based on the EM estimates. A practical implementation of the EM algorithm to the *P vivax* malaria progression is discussed in Section 3, compared to the currently existing two-stage method. The relaxed assumption for the transition probabilities of the baseline covariates is also discussed in Section 3. A simulation study in Section 4 presents the estimation results by the EM algorithm under different simulation scenarios. We analyze two *P vivax* infection datasets in Section 5 and show that our proposed method is feasible for practical use. Possible generalizations of our work are discussed in Section 6.

2 | MODEL AND ESTIMATION

2.1 | Model

For subject i, let X_i and Z_i denote the same covariates observed at the first observed disease occurrence (baseline) and recurrence, respectively. Let Y_i denote the disease outcome that follows a multinominal distribution with probability $\pi_{ik} = P(Y_i = k)$, $k = 0, 1, \dots, K$, and $\sum_{k=0}^{K}\pi_{ik} = 1$. We let $Y_i = 0$ indicate the subject is free of disease recurrence, and let $Y_i = k$ indicate the disease outcome is k, $k = 1, \dots, K$. Suppose that $X_i = (X_{i1}, \dots, X_{iJ})'$ is a J-column vector. Given a realization of $x_i = (x_{i1}, \dots, x_{iJ})'$, one can assume Y_i follows a multinomial logit model written by

$$\log\left\{\frac{\pi_{ik}(\theta_y)}{\pi_{i0}(\theta_y)}\right\} = \alpha_k + \beta_k' x_i, \tag{1}$$

for $k = 1, \dots, K$, where $\theta_y = (\alpha', \beta')'$, $\alpha = (\alpha_1, \dots, \alpha_K)'$, and $\beta = (\beta_1', \dots, \beta_K')'$ are parameters of interest with $\beta_k = (\beta_{k1}, \dots, \beta_{kJ})'$.

Model (1) holds for the association between the baseline covariate X_i and disease outcome Y_i. However, identifiability is an issue when Y_i is unknown if $Y_i > 0$. The latent outcome of the disease is interchangeable, and any permutation of the latent disease outcome has the same likelihood function. Therefore, one may not estimate β_k when Y_i is unknown. For the sake of identifiability, one may restrict the model to a more parsimonious one that concisely explains the covariate's association with the disease. For example, in malaria research, subjects who live in the epidemic area can be bitten by mosquitoes entirely at random. Hence, one may assume the new infection rate is constant and independent of X_i, that is, $\log\{\pi_{i1}(\theta)/\pi_{i0}(\theta)\} = \alpha_1$. However, the new infection rate may depend on genetic or biological factors.[22,23] One may build a regression model relating the new infection to those risk factors. Our approach still applies after such an adjustment.

When the disease recurs, one observes y_i as a realization of Y_i and $z_i = (z_{i1}, \dots, z_{iJ})'$ as a realization of Z_i that may have a different value from x_i. To model Z_i when $y_i > 0$, we assume the probability density function of Z_i possibly depends on x_i and y_i and has the form

$$f(z_i|x_i, y_i = k, \theta_z) = \prod_{j=1}^{J}\exp\left[\left\{z_{ij}g\left(\mu_{ijk}^z\right) - b\left(\mu_{ijk}^z\right)\right\}/a(\phi_{jk}) + c(z_{ij}, \phi_{jk})\right],$$

where $a(\cdot)$, $b(\cdot)$, $c(\cdot,\cdot)$ are known functions, ϕ_{jk} is the dispersion parameter, $g\left(\mu_{ijk}^z\right)$ is the canonical link function with mean $\mu_{ijk}^z = E(Z_{ij}|x_i, y_i = k)$, and θ_z contains every parameter in ϕ_{jk} and μ_{ijk}^z. Lin et al[20] termed the conditional density as the *transition likelihood*, describing the transition mechanism from X_i to Z_i.

The transition likelihood depends on the covariate type, which can be binary, normal, and Poisson. One could construct the transition likelihood function by applying the covariate's corresponding density function and link function. For example, when x_{ij} and z_{ij} are both binary, one can have the transition likelihood function written as $f(z_i|x_i, y_i = k, \theta_z) = \prod_{j=1}^{J}\exp\left[z_{ij}g\left(\mu_{ijk}^z\right) + \log\left(1 - \mu_{ijk}^z\right)\right]$ with logit link function $g\left(\mu_{ijk}^z\right) = \log\left\{\mu_{ijk}^z/\left(1 - \mu_{ijk}^z\right)\right\}$ and $\mu_{ijk}^z = P\left(Z_{ij} = 1\middle|x_{ij}, y_i = k, \theta_z\right) = x_{ij}q_{ijk} + (1 - x_{ij})\left(1 - q_{ijk}^*\right)$ with transition probabilities q_{ij} and q_{ij}^* for $x_{ij} = 1$ and $x_{ij} = 0$, respectively, for disease outcome k. The parameter θ_z includes corresponding parameters for q_{ijk} and q_{ijk}^*.

Note that, if the disease recurs, one observes $y_i > 0$. If the subject is disease-free, we assume $y_i = 0$ and Z_i has a point mass at x_i with probability density function $f(z_i|x_i, y_i = 0) = 1$ with no new information collected for the inference on the parameters.

2.2 | Estimation using EM algorithm

Let $\theta = (\theta_y', \theta_z')'$ be the parameter of interest that includes all of the model parameters for Y_i and Z_i. When the data $\mathcal{F}_i = \{(x_i, z_i, y_i)\}$ is fully observed for subject i, the full likelihood can be written as

$$L(\theta|x_i, z_i, y_i) = f(z_i|x_i, y_i, \theta) f(y_i|x_i, \theta) f(x_i|\theta). \tag{2}$$

Assuming x_i is fixed and contains no information on θ, the log-likelihood function is proportional to

$$\ell(\theta|x_i, z_i, y_i) = \log f(z_i|x_i, y_i, \theta) + \log f(y_i|x_i, \theta),$$

with

$$\log f(z_i|x_i, y_i, \theta) = \sum_{k=0}^{K} I(y_i = k) \log f(z_i|x_i, y_i = k, \theta),$$

and

$$\log f(y_i|x_i, \theta) = \sum_{k=0}^{K} I(y_i = k) \log\{\pi_{ik}(\theta)\},$$

where $\pi_{ik}(\theta) = P(Y_i = k|x_i, \theta)$ is the probability model of Y_i that follows model (1).

However, when y_i is missing, the observed data are $\mathcal{O}_i = \{x_i, \delta_i, \delta_i z_i\}$, where $\delta_i = I(Y_i > 0)$ indicates whether the disease progresses to any of the disease outcome $k = 1, \ldots, K$. Let $\mathcal{F} = \bigcup_{i=1}^{n} \mathcal{F}_i$ and $\mathcal{O} = \bigcup_{i=1}^{n} \mathcal{O}_i$ be data collected from n independent subjects with an identical distribution. One can write the Q function in the E-step of the EM algorithm by

$$\begin{aligned} Q(\theta|\theta_{\text{old}}) &= E\{\ell(\theta|\mathcal{F})|\mathcal{O}, \theta_{\text{old}}\} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_i P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}}) \log f(z_i|x_i, y_i = k, \theta) + \sum_{i=1}^{n} (1 - \delta_i) \log\{\pi_{i0}(\theta)\} \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} \delta_i P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}}) \log\{\pi_{ik}(\theta)\}, \end{aligned}$$

where

$$P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}}) = \frac{f(z_i|x_i, y_i = k, \theta_{\text{old}}) \pi_{ik}(\theta_{\text{old}})}{\sum_{k'=1}^{K} f(z_i|x_i, y_i = k', \theta_{\text{old}}) \pi_{ik'}(\theta_{\text{old}})}.$$

In the subsequent M-step, we aim to find the value of θ that maximizes the Q function with a starting value of θ_{old}. One can solve for a new θ by

$$\theta_{\text{new}} = \text{argmax}_{\theta} Q(\theta|\theta_{\text{old}}). \tag{3}$$

By making $\theta_{\text{old}} = \theta_{\text{new}}$ in the Q function and again solving the maximizer, one can obtain the updated estimate of θ. Repeating the E- and M-steps, one can find the estimate $\hat{\theta}$ that satisfies a predetermined convergence criterion.

2.3 | Classification

The primary aim of this article is to classify y_i into one of the disease categories when $\delta_i = 1$. With $\hat{\theta}$ denoted as the convergent estimate of θ in the EM algorithm, one can classify the unknown y_i with the largest $\pi_{ik}(\hat{\theta}) = P(Y_i = k|x_i, \hat{\theta})$ among $k = 1, \ldots, K$. However, Lin et al[20] have shown that this classification criterion may perform poorly since the classification uses only the baseline covariate information. To improve the classification performance, one can incorporate the covariate information of z_i observed when the disease recurs. As one can see, in the derivation for the EM algorithm, the classification probability $\pi_{ik}(\hat{\theta})$ can be updated using the likelihood function $f(z_i|x_i, y_i = k, \theta)$ under disease category $k = 1, \ldots, K$, namely,

$$P(Y_i = k|\mathcal{O}_i, \hat{\theta}) = \frac{f(z_i|x_i, y_i = k, \hat{\theta})\pi_{ik}(\hat{\theta})}{\sum_{k'=1}^{K} f(z_i|x_i, y_i = k', \hat{\theta})\pi_{ik'}(\hat{\theta})},$$

which can be considered a posterior probability that updates the prior probability $\pi_{ik}(\hat{\theta})$ using newly observed information z_i.

Given the observed information $\mathcal{O}_i$, one can classify the unknown y_i with the highest $P(Y_i = k|\mathcal{O}_i, \hat{\theta})$ among $k = 1, \ldots, K$. As discussed in Lin et al,[20] one can treat the classification probability $P(Y_i = k|\mathcal{O}_i, \theta)$ as an updated version of $\pi_{ik}(\theta)$ by a ratio of two transition likelihoods. One can show

$$\frac{P(Y_i = k|\mathcal{O}_i, \theta)}{P(Y_i = k'|\mathcal{O}_i, \theta)} = \frac{f(z_i|x_i, y_i = k, \theta)}{f(z_i|x_i, y_i = k', \theta)} \frac{\pi_{ik}(\theta)}{\pi_{ik'}(\theta)},$$

for any $k, k' \in \{1, \ldots, K\}$, which indicates the odds of the disease outcome k to k', in the sense of the prior probability $\pi_{ik}(\theta)$ to $\pi_{ik'}(\theta)$, is further updated by the ratio of the two transition likelihoods. This updating scheme guarantees a better, at least not worse, classification probability than $\pi_{ik}(\theta)$. The improvement depends on whether the conditional distribution of z_i given x_i is informative on the disease outcome y_i.

Since $\hat{\theta}$ is derived from the EM algorithm, statistical inferences on θ are possible and can be made by deriving the variance estimation through the Q function. One can show that $\sqrt{n}(\hat{\theta} - \theta)$ converges to a normal random variable with mean 0 and variance $I^{-1}(\theta)$, where $I(\theta) = \lim_{n\to\infty} n^{-1} E\{-\partial^2 \ell(\theta|\mathcal{O})/\partial\theta^2\}$. Empirically, $I(\theta)$ can be consistently estimated by $\hat{\sigma}^2 = -\partial^2 Q(\hat{\theta}|\hat{\theta})/\partial\theta^2$. However, as Louis[24] suggested, a better variance estimator is

$$\tilde{\sigma}^2 = \hat{\sigma}^2 - \sum_{i=1}^{n}\sum_{k=0}^{K} \omega_{ik}(\hat{\theta}) S_{ik}(\hat{\theta}) S_{ik}(\hat{\theta})' + \sum_{i=1}^{n} \left(\frac{\partial Q_i(\hat{\theta}|\hat{\theta})}{\partial\theta}\right)\left(\frac{\partial Q_i(\hat{\theta}|\hat{\theta})}{\partial\theta}\right)',$$

where $\omega_{ik}(\hat{\theta}) = (1 - \delta_i) + \delta_i P(Y_i = k|\mathcal{O}_i, \hat{\theta})$, $Q_i(\hat{\theta}|\hat{\theta}) = \sum_{k=0}^{K} \omega_{ik}(\hat{\theta}) \ell_{ik}(\hat{\theta})$ with $\ell_{ik}(\hat{\theta}) = \log f(z_i|x_i, y_i = k, \hat{\theta}) + \log\{\pi_{ik}(\hat{\theta})\}$, and $S_{ik}(\hat{\theta}) = \partial \ell_{ik}(\theta)/\partial\theta|_{\theta=\hat{\theta}}$.

3 | AN APPLICATION

3.1 | *P vivax* malaria progression

Lin et al[20] considered a *P vivax* malaria progression model $P(Y_i = k) = \pi_{ik}$ for $k = 0, 1, 2$, written as

$$\log\left\{\frac{\pi_{i1}(\theta)}{\pi_{i0}(\theta)}\right\} = \mu, \tag{4}$$

$$\log\left\{\frac{\pi_{i2}(\theta)}{\pi_{i0}(\theta)}\right\} = \alpha + \beta' x_i, \tag{5}$$

where $k = 0, 1, 2$ indicates no malaria recurrence ($k = 0$), new infection ($k = 1$), and relapse from the previous infection ($k = 2$), respectively. In their model, the new infection rate μ is a constant and $\theta = (\alpha, \beta')'$. A binary sequencing variant $x_i = (x_{i1}, \ldots, x_{iJ})'$ at baseline is assumed to be associated with the relapse in (5). They considered a transition model for

the recurrent variant Z_i in the case of relapse from the previous infection, written as

$$f(z_i|x_i, y_i = 2, \theta) = \prod_{j=1}^{J} \eta(x_{ij})^{z_{ij}} \{1 - \eta(x_{ij})\}^{1-z_{ij}}, \tag{6}$$

where

$$\eta(x_{ij}) = P(Z_{ij} = 1|x_{ij}, y_i = 2, \theta) = x_{ij}q_{ij} + (1 - x_{ij})(1 - q^*_{ij}),$$

with transition probabilities

$$q_{ij} = P(Z_{ij} = 1|x_{ij} = 1, y_i = 2),$$

and

$$q^*_{ij} = P(Z_{ij} = 0|x_{ij} = 0, y_i = 2).$$

For the reinfection, they assumed Z_i is independent of X_i but follows the same distribution written as

$$f(z_i|x_i, y_i = 1, \theta) = \prod_{j=1}^{J} p_j^{z_{ij}} (1 - p_j)^{1-z_{ij}}, \tag{7}$$

where $p_j = P(X_{ij} = 1)$ for $j = 1, \ldots, J$, indicating the prevalence of variant j.

Note that the estimation of q_{ij} and q^*_{ij} depends on the number of subjects with recurrent infections. When the number of covariates J is greater than the number of subjects with a recurrent infection, solutions of q_{ij} and q^*_{ij} may not be well defined. To avoid this issue, one may assume the transition probabilities q_{ij} and q^*_{ij} are the same for every subject i and covariate j, that is, $q_{ij} = q$ and $q^*_{ij} = q^*$ for every i and j. One may also assume the transition probability depends on a subject-level covariate w_{ij} via logistic models that can be written as

$$\log\left(\frac{q_{ij}}{1 - q_{ij}}\right) = \gamma_0 + \gamma_1 w_{ij}, \tag{8}$$

and

$$\log\left(\frac{q^*_{ij}}{1 - q^*_{ij}}\right) = \gamma^*_0, \tag{9}$$

where γ_1 is regarded as the association between w_{ij} and the likelihood of observing persistent variant j in the recurrent sequencing ($z_{ij} = 1$), given that the variant is observed in the baseline sequencing ($x_{ij} = 1$). The transition model using identical transition probabilities q and q^* is a special case of models (8) and (9). Here, we assume w_{ij} is observed only when the variant j is observed ($x_{ij} = 1$) and appears in the model (8). In the real data analysis, we use the reading frequency of the variant at the baseline sequencing as the covariate w_{ij}. General use of the subject-level covariate is possible. In that case, both models (8) and (9) can have the covariate as a predictor for the transition probabilities.

Note that one may not estimate μ and α in models (4) and (5) simultaneously since both parameters are part of the baseline relative recurrence of the infection. Lin et al[20] assumed μ is known or estimated via external information to avoid the identifiability problem. We use the same strategy in our real data analysis.

3.2 | EM algorithm for the application

In this application, we observe $\delta_i = I(y_i > 0)$, meaning whether the patient has recurrent infection, that is, $\delta_i = 1$ when $y_i = 1, 2$ (reinfection or relapse) and $\delta_i = 0$ when $y_i = 0$ (no recurrent infection). Since the covariate is collected when

the recurrent infection occurs, we let $f(z_i|x_i, y_i = 0, \theta) = 1$, meaning Z_i is not a random variable when the subject is infection-free. Therefore, the Q function can be simplified to

$$Q(\theta|\theta_{\text{old}}) = Q_1(\theta|\theta_{\text{old}}) + Q_2(\theta|\theta_{\text{old}}),$$

where

$$Q_1(\theta|\theta_{\text{old}}) = \sum_{i=1}^{n}(1-\delta_i)\log\{\pi_{i0}(\theta)\} + \sum_{i=1}^{n}\sum_{k=1}^{2}\delta_i P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}})\log\{\pi_{ik}(\theta)\},$$

and

$$Q_2(\theta|\theta_{\text{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{2}\delta_i P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}})\log f(z_i|x_i, y_i = k, \theta),$$

with

$$P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}}) = \frac{f(z_i|x_i, y_i = k, \theta_{\text{old}})\pi_{ik}(\theta_{\text{old}})}{\sum_{\ell=1}^{2} f(z_i|x_i, y_i = \ell, \theta_{\text{old}})\pi_{i\ell}(\theta_{\text{old}})}.$$

Iteratively maximizing $Q(\theta|\theta_{\text{old}})$, we can obtain the maximum likelihood estimate $\hat{\theta}$ that maximizes the observed likelihood function.

We apply models (4) and (5) for $\pi_{ik}(\theta)$, $k = 0, 1, 2$, in $Q_1(\theta|\theta_{\text{old}})$, where

$$\pi_{i1}(\theta) = \frac{\exp(\mu)}{1 + \exp(\mu) + \exp(\alpha + \beta' x_i)},$$
$$\pi_{i2}(\theta) = \frac{\exp(\alpha + \beta' x_i)}{1 + \exp(\mu) + \exp(\alpha + \beta' x_i)},$$

and $\pi_{i0}(\theta) = 1 - \pi_{i1}(\theta) - \pi_{i2}(\theta)$. We apply models (6) to (9) for the transition likelihood functions $f(z_i|x_i, y_i = 1, \theta)$ and $f(z_i|x_i, y_i = 2, \theta)$. The unknown parameters may include the prevalence $p_j = P(X_{ij} = 1)$. However, since x_{ij} is always observed, one can consistently estimate p_j by $\hat{p}_j = n^{-1}\sum_{i=1}^{n} x_{ij}$. To reduce the number of parameters for faster convergence of estimates, we replace p_j with $\hat{p}_j$ in θ_2. Accordingly, we implement the EM algorithm to solve for $\theta = (\alpha, \beta', \gamma', \gamma_0^*)'$, where $\gamma' = (\gamma_0, \gamma_1)'$.

With $\hat{\theta}$ as the convergent estimate of θ using the EM algorithm, we calculate the classification probability

$$P(Y_i = k|\mathcal{O}_i, \hat{\theta}) = \frac{f(z_i|x_i, y_i = k, \hat{\theta})\pi_{ik}(\hat{\theta})}{\sum_{\ell=1}^{2} f(z_i|x_i, y_i = \ell, \hat{\theta})\pi_{i\ell}(\hat{\theta})}, \tag{10}$$

for each subject i and classify the recurrent infection as relapse if $P(Y_i = 2|\mathcal{O}_i, \hat{\theta}) > P(Y_i = 1|\mathcal{O}_i, \hat{\theta})$, or simply $P(Y_i = 2|\mathcal{O}_i, \hat{\theta}) > 0.5$.

3.3 | Comparison with previous work

Based on the observed data $O = \bigcup_{i=1}^{n} O_i$, where $O_i = \{x_i, \delta_i, \delta_i z_i\}$, Lin et al[20] utilized the observed log-likelihood function, $\ell(\theta) = \sum_{i=1}^{n}\log f(z_i|x_i, \delta_i = 1, \theta) + \sum_{i=1}^{n}\log f(\delta_i|x_i, \theta) + \sum_{i=1}^{n}\log f(x_i|\theta)$, and proposed a two-stage method that first used the baseline sequencing variant x_i to obtain the estimator $\hat{\theta}_1^*$ of $\theta_1 = (\alpha, \beta')'$ in the model (5). The target function they maximized is

$$\ell_1(\theta_1) = \sum_{i=1}^{n}\log f(\delta_i|x_i, \theta_1) = \sum_{i=1}^{n}(1-\delta_i)\log\{\pi_{i0}(\theta_1)\} + \sum_{i=1}^{n}\delta_i\log\{1 - \pi_{i0}(\theta_1)\},$$

which is different from the Q_1 function that utilizes posterior probability $P(Y_i = k|\mathcal{O}_i, \theta_{\text{old}})$. Instead, the second term of the $\ell_1(\theta_1)$ function uses the observed outcome $\delta_i = I(Y_i > 0)$ with probability $P(\delta_i = 1) = \pi_{i1}(\theta_1) + \pi_{i2}(\theta_1) = 1 - \pi_{i0}(\theta_1)$.

In the second stage, when using both baseline and recurrent sequencing results, they maximized

$$\ell_2(\theta_2) = \sum_{i=1}^{n} \delta_i (1 - \hat{\xi}_i^{(0)}) \log f(z_i|x_i, y_i = 1, \hat{\theta}_1^*, \theta_2) + \sum_{i=1}^{n} \delta_i \hat{\xi}_i^{(0)} \log f(z_i|x_i, y_i = 2, \hat{\theta}_1^*, \theta_2),$$

where $\hat{\theta}_1^* = (\hat{\alpha}, \hat{\beta}')'$, $\theta_2 = (\gamma', \gamma^*)'$, and $\hat{\xi}_i^{(0)} = \pi_{i2}(\hat{\theta}_1^*)/\{\pi_{i1}(\hat{\theta}_1^*) + \pi_{i2}(\hat{\theta}_1^*)\}$ is the classification probability obtained from the first stage.

The target function $\ell_2(\theta_2)$ is derived from the mixture distribution of z_i with mixture probability replaced by the estimates $\hat{\xi}_i^{(0)}$ from the first stage. It differs from the Q_2 function that uses the posterior probability $P(Y_i = 2|\mathcal{O}_i, \theta_{\text{old}})$ as the mixture probability. In fact, with $\hat{\theta}_2^*$ denoted for the maximizer of θ_2 in $\ell_2(\theta_2)$, the classification probability $\hat{\xi}_i^{(1)}$ proposed in Lin et al[20] can be written as

$$\hat{\xi}_i^{(1)} = \frac{f(z_i|x_i, y_i = 2, \hat{\theta}_1^*, \hat{\theta}_2^*)\pi_{ik}(\hat{\theta}_1^*)}{\sum_{\ell=1}^{2} f(z_i|x_i, y_i = \ell, \hat{\theta}_1^*, \hat{\theta}_2^*)\pi_{i\ell}(\hat{\theta}_1^*)},$$

which can be considered as $P(Y_i = 2|\mathcal{O}_i, \hat{\theta}^*)$ with $\hat{\theta}^* = (\hat{\theta}_1^{*\prime}, \hat{\theta}_2^{*\prime})'$ and compared with $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ in (10). We compare the performance of the two classification probabilities $\hat{\xi}_i^{(1)}$ and $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ in both simulation studies and real data analysis.

To evaluate the precision of estimators $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ under the two-stage method, we estimate the variances of the estimators using bootstrap. First, one can obtain a bootstrapped sample with n observations by drawing $(x_i^*, \delta_i^*, \delta_i^* z_i^*)$ with replacement from the original data $(x_i, \delta_i, \delta_i z_i)$, $i = 1, \ldots, n$. Then, one can obtain $\hat{\theta}_1^{*(1)}$ and $\hat{\theta}_2^{*(1)}$ using the two-stage method based on the bootstrapped sample $(x_i^*, \delta_i^*, \delta_i^* z_i^*)$. By repeating the resampling and estimation steps B times, one can obtain $\hat{\theta}_1^{*(1)}, \ldots, \hat{\theta}_1^{*(B)}$ and $\hat{\theta}_2^{*(1)}, \ldots, \hat{\theta}_2^{*(B)}$. The variance of $\hat{\theta}_1^*$ can be estimated by $\hat{\sigma}_B^2(\hat{\theta}_1^*) = (B-1)^{-1}\sum_{b=1}^{B}(\hat{\theta}_1^{*(b)} - \overline{\theta}_1^*)^2$, where $\overline{\theta}_1^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}_1^{*(b)}$. The variance of $\hat{\theta}_2^*$ can be estimated similarly. One can simply calculate the standard deviation of the replications for the bootstrapped standard error of the estimators $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$.

4 | SIMULATION STUDY

In this section, we demonstrate our methodology via comprehensive simulation experiments. We mimic the *P vivax* malaria infection data with three disease outcomes, namely, no recurrent infection ($y = 0$), new infection ($y = 1$), and relapse from the previous infection ($y = 2$). We assume the relative infection rate $\mu = -2, -3$ in the model (4) for the new infection, and $\alpha = -2$ and $\beta = (\log(2), \log(2), \log(2), 0, \ldots, 0)'$ in the ten variants model (5) for the relapse, meaning three most prevalent variants are associated with the occurrence of relapse. We generated x_i following Bernoulli distribution with probability $P(X_j = 1) = 0.5\exp\{-0.1(j-1)\}$ for $j = 1, \ldots, 10$. We generated z_i following the transition model (6) if $y_i = 2$ and model (7) if $y_i = 1$. We evaluate the performance of the EM algorithm and two-stage method in two scenarios. In scenario 1, we assume the transition probabilities are the same for each variant, that is, $q_1 = q_2 = \cdots = q_J$ and $q_1^* = q_2^* = \cdots = q_J^*$, and the transition probabilities follow logistic models (8) and (9) with $\gamma_1 = 0$. There are only two parameters γ_0 and γ_0^* in this case. We let $\gamma_0^* = 2.94$ ($q^* = 0.95$) and $\gamma_0 = 0, 2.94$ ($q = 0.5, 0.95$), where a smaller value of γ_0 makes model (6) similar to model (7), that is, the transition signal from $y_i = 2$ is weaker when γ_0 is smaller. One can anticipate that the classifier using the transition likelihood improves less when $\gamma_0 = 0$. In scenario 2, we assume the logistic models include the covariate ω_{ij} with $\gamma_1 = 0$ or $\gamma_1 = 0.5$. We generate ω_{ij} uniformly between 0 and 1. The sample size takes three different values, $n = 100, 200, 400$, and 1000 repetitions were made for each combination of μ, γ_0, and n in each scenario.

Table 1 shows the operating characteristics of classifiers by the two-stage method and EM algorithm. From Table 1, one can see that both classifiers perform aggressively under a low transition probability when the recurrent infection is a relapse ($y = 2$). Most recurrences were claimed as relapses, resulting in high sensitivity but low specificity. Both classifiers perform well under a high transition probability, reaching a high degree of accuracy in both sensitivity and specificity. When comparing the two classifiers $\hat{\xi}_i^{(1)}$ and $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$, one can see that $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ has higher accuracy than

TABLE 1 Operating characteristics of the classifiers by the two-stage method and the EM algorithm

			Two-stage method $\hat{\xi}_i^{(1)} > 0.5$			EM algorithm $P(Y_i = 2\|\mathcal{O}_i, \hat{\theta}) > 0.5$		
μ	γ_0	n	Sensitivity	Specificity	Overall	Sensitivity	Specificity	Overall
−3	0	100	96.5%	53.1%	91.4%	97.4%	56.7%	92.7%
		200	98.3%	57.0%	93.4%	98.1%	59.2%	93.5%
		400	98.4%	58.3%	93.7%	98.3%	59.3%	93.7%
	2.94	100	97.8%	83.5%	96.1%	98.9%	85.6%	97.4%
		200	99.2%	86.3%	97.6%	99.1%	86.9%	97.7%
		400	99.2%	86.2%	97.7%	99.2%	86.5%	97.7%
−2	0	100	84.0%	64.2%	78.5%	93.4%	69.7%	87.0%
		200	94.0%	69.5%	87.4%	94.7%	71.4%	88.6%
		400	95.5%	71.0%	88.9%	95.3%	72.0%	89.1%
	2.94	100	87.5%	85.0%	86.7%	97.6%	89.9%	95.6%
		200	97.4%	90.4%	95.6%	98.1%	90.6%	96.2%
		400	98.2%	91.3%	96.4%	98.3%	91.6%	96.5%

$\hat{\xi}_i^{(1)}$, especially when the sample size n is small. When the sample size increases, the two-stage method $\hat{\xi}_i^{(1)}$ becomes more competitive and has a similar classification performance. Note that the two-stage method enjoys the advantage of fast computing. One can use the estimate of the two-stage method as the starting point θ_{old} for the EM algorithm. An out-of-sample prediction comparison using ten-fold cross-validation is shown in the supplementary material. The result shows that our proposed classifier using the EM algorithm outperforms the two-stage method in prediction.

Table 2 shows the simulation results for estimating regression coefficients and transition probabilities using the EM algorithm when the transition probabilities under a relapse are the same for each variant ($\gamma_1 = 0$). We report bias (b), empirical standard error (σ), estimated standard error ($\hat{\sigma}$), and coverage probabilities (CP) for the regression coefficients and transition probabilities. To save space, we only show the estimation results for β_1 to β_5 and γ_0 and γ_0^* in Table 2. Other parameter estimations are shown in the supplementary material.

Table 3 shows the simulation results for estimating regression coefficients and transition probabilities by the two-stage method with variance estimation using $B = 100$ bootstrapped samples. From Table 3, one can see the bootstrapped standard errors are generally larger than the empirical standard errors, resulting in high coverage probabilities. When comparing Tables 2 and 3, one can see the EM algorithm has a better performance overall and a better standard error estimation than the bootstrapped estimation, especially when the sample size is small. When the sample size increases, the two-stage method with bootstrapped variance estimation performs better but still has a conservative variance estimation. The coverage probability is generally higher than the 0.95 nominal level.

Tables 4 and 5 show the estimation of regression coefficients and transition probabilities using the EM algorithm at $\gamma_1 = 0$ and $\gamma_1 = 0.5$, respectively, when the transition probability follows a logistic model. From Tables 4 and 5, one can observe that the EM algorithm's performance is satisfactory when transition probabilities follow logistic models. This result led us to put our insight into the computation under high dimensionality in the future.

5 | REAL DATA ANALYSIS

5.1 | *P vivax* infection in northern Cambodia

It is well-known that many clones and strains exist within a *P vivax* infected human host.[25] In a treatment study conducted in northern Cambodia from 2010 to 2011, the *P vivax* merozoite surface protein 1 (*pvmsp1*) gene was found to have great nucleotide diversity,[25,26] making targeted amplicon deep sequencing an excellent tool to genotype isolates from *P vivax* infected patients.[19] The *pvmsp1* sequence variants were determined by a bioinformatics pipeline using a clustering method to construct the most likely haplotypes within a patient.[27] As a result, 67 unique *pvmsp1* haplotypes (variants)

TABLE 2 The bias (b), empirical standard error (σ), estimated standard error ($\hat{\sigma}$), and coverage probability (CP) for the regression coefficient estimation using EM algorithms for scenario 1 when the transition probabilities under a relapse are the same for each variant

μ	γ_0	n	b_1	b_2	b_3	b_4	b_5	b_{γ_0}	$b_{\gamma_0^*}$	σ_1	σ_2	σ_3	σ_4	σ_5	σ_{γ_0}	$\sigma_{\gamma_0^*}$
−3	0	100	0.14	0.17	0.11	0.01	−0.03	−0.01	0.15	0.70	0.67	0.68	0.68	0.83	0.22	1.21
		200	0.05	0.06	0.04	0.02	0.01	−0.01	0.02	0.41	0.38	0.39	0.39	0.39	0.15	0.32
		400	0.02	0.02	0.02	0.00	0.00	0.00	0.01	0.27	0.25	0.26	0.26	0.27	0.10	0.21
	2.94	100	0.12	0.15	0.10	0.00	−0.01	0.40	0.08	0.64	0.63	0.63	0.64	0.65	2.10	0.78
		200	0.05	0.06	0.03	0.02	0.01	0.06	0.03	0.39	0.37	0.38	0.38	0.37	0.41	0.28
		400	0.02	0.03	0.02	0.01	0.00	0.01	0.02	0.26	0.24	0.25	0.26	0.26	0.26	0.19
−2	0	100	0.21	0.23	0.23	−0.07	−0.04	0.02	0.28	1.12	1.53	1.52	0.94	1.09	0.24	1.48
		200	0.05	0.07	0.07	−0.03	0.00	0.01	0.06	0.43	0.44	0.42	0.45	0.46	0.16	0.39
		400	0.03	0.04	0.04	−0.02	0.00	0.01	0.02	0.27	0.28	0.27	0.29	0.31	0.11	0.25
	2.94	100	0.15	0.15	0.14	−0.03	−0.01	0.52	0.13	0.68	0.67	0.68	0.71	0.67	2.32	0.80
		200	0.05	0.06	0.06	−0.02	0.00	0.09	0.04	0.39	0.40	0.39	0.41	0.43	0.58	0.29
		400	0.02	0.03	0.04	−0.01	0.00	0.03	0.02	0.25	0.26	0.25	0.27	0.28	0.28	0.21

μ	γ_0	n	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_5$	$\hat{\sigma}_{\gamma_0}$	$\hat{\sigma}_{\gamma_0^*}$	CP$_1$	CP$_2$	CP$_3$	CP$_4$	CP$_5$	CP$_{\gamma_0}$	CP$_{\gamma_0^*}$
−3	0	100	0.56	0.56	0.56	0.57	0.59	0.20	0.40	94.1	94.4	92.9	93.4	93.4	94.2	93.9
		200	0.37	0.37	0.37	0.38	0.39	0.14	0.30	93.9	95.0	94.2	94.2	95.7	93.9	94.9
		400	0.25	0.25	0.25	0.26	0.26	0.10	0.20	94.1	95.4	93.5	94.5	95.4	94.1	94.6
	2.94	100	0.55	0.55	0.54	0.56	0.57	0.50	0.37	94.1	94.6	94.2	94.5	94.3	95.2	93.6
		200	0.36	0.36	0.36	0.37	0.38	0.37	0.27	94.0	95.2	94.4	94.4	95.6	95.2	94.6
		400	0.25	0.24	0.24	0.25	0.26	0.25	0.18	94.2	95.9	94.4	95.2	94.0	94.0	94.9
−2	0	100	0.62	0.62	0.62	0.64	0.65	0.21	0.46	93.2	92.8	93.3	91.8	94.4	93.1	92.7
		200	0.41	0.41	0.41	0.42	0.43	0.15	0.35	93.7	94.0	95.4	94.3	94.6	93.6	94.1
		400	0.27	0.27	0.27	0.28	0.28	0.11	0.23	95.5	95.7	95.0	94.8	93.7	94.2	92.8
	2.94	100	0.57	0.57	0.57	0.58	0.59	0.53	0.39	93.3	92.6	94.1	92.9	95.6	95.0	95.7
		200	0.37	0.37	0.37	0.38	0.39	0.37	0.27	94.5	94.2	94.8	94.1	93.3	95.4	94.8
		400	0.25	0.25	0.25	0.26	0.26	0.26	0.19	95.7	95.1	94.9	94.9	93.9	94.5	94.9

Note: The estimates are computed over 1000 repetitions of sample size $n = 100$, 200, and 400 for each combination of the new infection rate μ, γ_0, and n.

were detected among 78 *P vivax*-infected subjects. Among them, nine haplotypes appeared in at least 10% of individuals. This analysis uses the nine most frequent variants as the model (5) covariates for model building and classification. During the follow-up period, 55 (71%) subjects were free of recurrent infection. Assuming the 5% reinfection rate suggested in Lin et al,[20] we used $\mu = -3$ in the model (4) since $\log(0.05/0.71) \approx -3$.

Table 6 shows the classification result based on the EM algorithm and the two-stage method for the 23 subjects with recurrent infection. We report baseline and recurrence variants, two classification probabilities, classification results based on the posterior probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ and two-stage method $\hat{\xi}_i^{(1)}$. We use their identification number in our data set to show the pair of infections from baseline to recurrence, for example, 10→10R, and use letters to represent the observed variants. For example, in pair 10→10R, we have $x_i = z_i = (1, 0, \ldots, 0)'$ since variant A was observed at both sequencing results. The recurrence sequencing for pairs 31 → 31R and 68 → 68R are blank because the observed variants are less prevalent. Most recurrences are classified as relapses because of a high degree of overlap in dominant variants. For pairs with reinfection as the classification result, they mostly have at least one non-sharing variant in the recurrence sequencing, such as pairs 118 → 118R, 151 → 151R, 160 → 160R, and 179 → 179R. Other pairs classified as reinfection have multiple non-sharing variants between baseline and recurrence sequencing, resulting in a low posterior classification probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$. For example, in the 125 → 125R pair, the variants A, B, and E that appeared in the recurrence

TABLE 3 The bias (b), empirical standard error (σ), estimated standard error ($\hat{\sigma}$), and coverage probability (CP) for the regression coefficient estimation using the two-stage method for scenario 1 when the transition probabilities under a relapse are the same for each variant

μ	γ_0	n	b_1	b_2	b_3	b_4	b_5	b_{γ_0}	$b_{\gamma_0^*}$	σ_1	σ_2	σ_3	σ_4	σ_5	σ_{γ_0}	$\sigma_{\gamma_0^*}$
−3	0	100	0.54	0.59	0.39	−0.06	−0.17	0.00	0.03	4.65	5.16	2.55	2.50	3.25	0.22	1.05
		200	0.06	0.05	0.05	−0.02	−0.01	0.00	0.00	0.41	0.41	0.43	0.44	0.44	0.15	0.31
		400	0.05	0.04	0.03	0.01	−0.02	0.00	0.01	0.28	0.28	0.28	0.27	0.28	0.10	0.22
	2.94	100	0.54	0.59	0.39	−0.06	−0.17	0.40	0.08	4.72	5.16	2.55	2.50	3.25	2.65	1.03
		200	0.06	0.05	0.05	−0.02	−0.01	0.07	0.02	0.41	0.41	0.43	0.44	0.44	0.76	0.28
		400	0.05	0.04	0.03	0.01	−0.02	0.02	0.01	0.28	0.28	0.28	0.27	0.28	0.26	0.19
−2	0	100	3.60	3.85	3.15	−0.42	−0.01	−0.04	−0.11	14.5	14.3	13.6	13.6	11.3	0.25	1.51
		200	0.17	0.15	0.12	−0.03	0.02	−0.01	−0.02	1.29	0.91	0.89	0.89	0.59	0.16	0.37
		400	0.04	0.04	0.05	−0.02	0.00	0.00	−0.03	0.34	0.33	0.33	0.34	0.34	0.11	0.24
	2.94	100	3.60	3.85	3.15	−1.12	−0.70	0.34	0.04	14.5	14.3	13.6	13.6	11.3	3.32	1.50
		200	0.17	0.15	0.12	−0.03	0.02	0.03	0.02	1.29	0.91	0.89	0.89	0.59	0.44	0.30
		400	0.04	0.04	0.05	−0.02	0.00	0.02	0.00	0.34	0.33	0.33	0.34	0.34	0.29	0.21
μ	γ_0	n	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_5$	$\hat{\sigma}_{\gamma_0}$	$\hat{\sigma}_{\gamma_0^*}$	CP_1	CP_2	CP_3	CP_4	CP_5	CP_{γ_0}	$CP_{\gamma_0^*}$
−3	0	100	13.6	13.1	13.0	12.5	13.8	0.68	3.78	99.6	99.6	99.7	99.8	99.9	99.2	98.3
		200	1.42	1.36	1.31	1.12	1.25	0.16	0.72	99.5	99.4	99.5	99.7	99.8	96.0	97.4
		400	0.30	0.29	0.29	0.30	0.30	0.11	0.24	96.2	96.6	97.0	97.5	96.6	96.6	95.8
	2.94	100	13.6	13.1	13.0	12.5	13.8	5.31	3.56	99.6	99.7	99.8	99.9	100	97.7	98.7
		200	1.42	1.36	1.31	1.12	1.25	1.23	0.44	99.5	99.4	99.5	99.7	99.8	97.0	97.0
		400	0.30	0.29	0.29	0.30	0.31	0.31	0.20	94.0	99.0	95.0	98.0	97.0	97.4	95.1
−2	0	100	20.1	19.6	19.1	19.9	20.6	1.06	4.13	97.9	97.7	97.4	98.4	99.0	99.5	98.1
		200	7.69	7.36	6.78	6.71	7.37	0.27	1.51	100	100	99.9	100	100	98.3	98.6
		400	0.59	0.54	0.52	0.47	0.51	0.12	0.32	98.5	98.7	98.5	98.3	98.2	96.7	97.1
	2.94	100	20.1	19.6	19.1	19.9	20.6	5.88	4.29	97.9	97.7	97.4	98.4	99.0	97.3	99.4
		200	7.69	7.36	6.78	6.71	7.37	2.39	1.12	100	100	99.9	100	100	99.3	98.9
		400	0.59	0.54	0.52	0.47	0.51	0.40	0.22	98.5	98.7	98.5	98.3	98.2	97.1	95.2

Note: The estimates are computed over 1000 repetitions of sample size $n = 100$, 200, and 400 for each combination of new infection rate μ, γ_0, and n.

sequencing are unobserved in the baseline sequencing, resulting in a low $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$. However, non-sharing variants in the baseline sequencing have little impact on the recurrence classification probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$. Taking pairs 36 → 36R and 126 → 126R for example, the classification probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ remains high, even when multiple variants were unobserved in the recurrence sequencing. It should be noted that compared to our method, the two-stage method can be conservative in some situations. Pairs like 80 → 80R, 152 → 152R and 154 → 154R that have both prevalent variants overlapping and non-sharing variant appeared in the recurrence sequencing are more likely classified reinfections by the two-stage method while classified as relapses by our method.

Table 7 shows the estimation of regression coefficients by the EM algorithm. We report variants' prevalence, regression coefficient estimate, standard error, and P-value by a Wald-type test. As one can see, variants A, C, E, and D with relatively high prevalence are not statistically significant. However, the association between variant A and relapse is large ($\hat{\beta} = 10.86$, $p = 0.561$), which is likely due to sparsity in the data. Variant E is the only variant reaching the statistical significance with a positive association ($\hat{\beta} = 0.346$, $p = 0.006$). To estimate the parameters related to transition probabilities, we have $\hat{\gamma}_1 = 0.01$, indicating the variant's frequency is positively associated with the transition of the variant in relapse. Also, parameter estimates $\hat{\gamma}_0 = -0.85$ and $\hat{\gamma}_0^* = 1.75$ show that a baseline variant in relapse is likely absent in the

TABLE 4 The bias (b), empirical standard error (σ), estimated standard error ($\hat{\sigma}$), and coverage probability (CP) for the regression coefficient estimation using EM algorithms for scenario 2 when the transition probability follows a logistic model with $\gamma_1 = 0$

μ	γ_0	n	b_1	b_2	b_3	b_4	b_{γ_0}	b_{γ_1}	$b_{\gamma_0^*}$	σ_1	σ_2	σ_3	σ_4	σ_{γ_0}	σ_{γ_1}	$\sigma_{\gamma_0^*}$
−3	0	100	0.16	0.19	0.08	−0.03	0.02	−0.04	0.12	0.70	1.05	0.66	0.70	0.44	0.76	0.87
		200	0.06	0.06	0.02	−0.02	0.00	0.00	0.03	0.39	0.39	0.39	0.39	0.31	0.53	0.32
		400	0.03	0.02	0.01	−0.01	−0.01	0.01	0.02	0.26	0.25	0.25	0.25	0.21	0.36	0.21
	2.94	100	0.14	0.14	0.06	−0.02	0.61	0.10	0.07	0.62	0.65	0.60	0.63	2.97	4.13	0.43
		200	0.06	0.06	0.02	−0.02	0.14	0.00	0.04	0.37	0.38	0.37	0.40	0.87	1.50	0.29
		400	0.03	0.02	0.01	−0.01	0.07	−0.01	0.02	0.26	0.25	0.25	0.25	0.54	0.94	0.20
−2	0	100	0.17	0.21	0.21	−0.08	−0.02	0.04	0.31	1.53	1.47	1.19	1.14	0.46	0.80	1.59
		200	0.04	0.05	0.06	−0.01	−0.02	0.03	0.07	0.43	0.41	0.43	0.45	0.32	0.55	0.52
		400	0.02	0.02	0.03	−0.01	−0.01	0.02	0.02	0.28	0.27	0.28	0.30	0.23	0.40	0.25
	2.94	100	0.12	0.12	0.15	−0.04	0.80	0.34	0.12	0.68	0.67	0.65	0.66	3.72	6.25	0.77
		200	0.04	0.04	0.06	−0.01	0.16	0.11	0.04	0.40	0.37	0.40	0.40	1.01	2.40	0.31
		400	0.02	0.02	0.02	−0.01	0.09	−0.03	0.01	0.26	0.25	0.27	0.27	0.61	1.02	0.21

μ	γ_0	n	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_{\gamma_0}$	$\hat{\sigma}_{\gamma_1}$	$\hat{\sigma}_{\gamma_0^*}$	CP_1	CP_2	CP_3	CP_4	CP_{γ_0}	CP_{γ_1}	$CP_{\gamma_0^*}$
−3	0	100	0.57	0.56	0.57	0.58	0.41	0.71	0.41	93.9	93.4	94.7	93.3	95.3	95.2	95.8
		200	0.37	0.36	0.37	0.38	0.29	0.49	0.29	94.4	93.7	95.3	95.7	92.7	94.4	94.4
		400	0.25	0.25	0.25	0.26	0.20	0.35	0.20	95.0	94.7	95.1	95.8	94.2	94.3	94.8
	2.94	100	0.55	0.54	0.55	0.56	1.00	1.82	0.37	94.2	93.4	95.0	94.2	95.3	97.0	95.0
		200	0.36	0.36	0.36	0.37	0.70	1.23	0.26	95.0	93.8	95.1	95.6	94.6	95.8	94.6
		400	0.25	0.24	0.24	0.25	0.49	0.85	0.18	94.8	95.3	95.9	95.6	95.8	95.1	94.2
−2	0	100	0.63	0.62	0.62	0.65	0.43	0.74	0.46	90.6	93.0	92.9	92.7	94.3	94.7	92.3
		200	0.40	0.40	0.40	0.41	0.30	0.52	0.32	93.6	95.3	93.9	94.2	94.3	94.0	93.6
		400	0.27	0.27	0.27	0.28	0.21	0.36	0.22	94.4	94.9	94.2	94.3	93.0	93.0	93.8
	2.94	100	0.57	0.57	0.57	0.58	1.04	1.94	0.39	92.8	93.8	93.9	93.6	94.7	96.7	93.6
		200	0.37	0.37	0.37	0.38	0.73	1.31	0.27	93.7	95.8	93.8	94.4	95.2	95.8	94.9
		400	0.25	0.25	0.25	0.26	0.52	0.90	0.19	95.4	95.0	94.0	93.9	94.5	94.7	94.5

Note: The estimates are computed over 1000 repetitions of sample size $n = 100$, 200, and 400 for each combination of new infection rate μ, γ_0, and n.

recurrence, and a variant absent in the baseline will likely remain absent in the recurrent infection. The corresponding standard errors for $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_0^*$ are 0.37, 0.01, and 0.31, respectively.

5.2 | *P vivax* infection in southern Cambodia

We demonstrate our methodology in another malaria treatment study conducted between August 2006 and February 2008 in Chumkiri District, Kampot Province, in southern Cambodia.[28] The study enrolled 110 subjects with uncomplicated *P vivax* malaria and treated them with artesunate-mefloquine therapy, which is highly effective for vivax blood-stage parasite, but does not kill hypnozoites in the liver (that can later emerge to cause relapse). Of the 107 subjects who completed the follow-up, 45 (42.1%) suffered recurrent *P vivax* parasitemia. All recurrent parasitaemias occurred between day 28 and day 42. No attempt was made to distinguish between recrudescence (treatment failure), relapse from the liver, and reinfection in the original trial. Since recrudescence after ACT therapy is unlikely for *P vivax*, we applied our method to classify the 45 recurrences as either relapse or reinfection. For this analysis, we used genotyping information for each baseline and recurrence derived from heteroduplex tracking assays targeting *Pvmsp1*.[29] In total, 16 unique *pvmsp1* variants were

TABLE 5 The bias (b), empirical standard error (σ), estimated standard error ($\hat{\sigma}$), and coverage probability (CP) for the regression coefficient estimation using EM algorithms for scenario 2 when the transition probability follows a logistic model with $\gamma_1 = 0.5$

μ	γ_0	n	b_1	b_2	b_3	b_4	b_{γ_0}	b_{γ_1}	$b_{\gamma_0^*}$	σ_1	σ_2	σ_3	σ_4	σ_{γ_0}	σ_{γ_1}	$\sigma_{\gamma_0^*}$
−3	0	100	0.15	0.18	0.08	−0.04	−0.01	0.03	0.10	0.67	0.97	0.64	0.70	0.43	0.77	0.69
		200	0.06	0.06	0.02	−0.02	0.00	0.01	0.03	0.38	0.39	0.39	0.39	0.31	0.53	0.32
		400	0.03	0.02	0.01	−0.01	0.00	0.00	0.02	0.26	0.25	0.25	0.25	0.21	0.37	0.21
	2.94	100	0.14	0.13	0.06	−0.02	0.83	0.45	0.07	0.61	0.64	0.60	0.63	3.67	5.63	0.43
		200	0.06	0.06	0.02	−0.01	0.17	0.09	0.04	0.37	0.37	0.38	0.37	1.18	1.82	0.29
		400	0.03	0.02	0.01	0.00	0.06	0.04	0.02	0.26	0.25	0.24	0.25	0.60	1.14	0.20
−2	0	100	0.15	0.17	0.20	−0.06	0.00	0.04	0.24	0.81	0.94	0.82	0.91	0.47	0.87	1.33
		200	0.05	0.05	0.06	−0.01	0.01	0.00	0.05	0.43	0.41	0.43	0.44	0.32	0.56	0.38
		400	0.02	0.02	0.03	−0.01	0.01	−0.01	0.01	0.28	0.27	0.28	0.29	0.22	0.39	0.25
	2.94	100	0.12	0.12	0.15	−0.04	1.03	0.57	0.11	0.68	0.66	0.65	0.66	4.23	6.95	0.69
		200	0.04	0.04	0.06	−0.01	0.24	0.21	0.04	0.40	0.37	0.40	0.40	1.61	3.21	0.31
		400	0.02	0.02	0.02	−0.01	0.11	0.01	0.01	0.26	0.25	0.27	0.27	0.70	1.27	0.21

μ	γ_0	n	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\sigma}_4$	$\hat{\sigma}_{\gamma_0}$	$\hat{\sigma}_{\gamma_1}$	$\hat{\sigma}_{\gamma_0^*}$	CP_1	CP_2	CP_3	CP_4	CP_{γ_0}	CP_{γ_1}	$CP_{\gamma_0^*}$
−3	0	100	0.57	0.56	0.57	0.58	0.41	0.72	0.41	94.9	93.9	94.9	94.0	94.9	95.7	95.1
		200	0.37	0.36	0.36	0.38	0.29	0.50	0.28	94.9	93.6	94.9	95.3	94.4	94.0	94.2
		400	0.25	0.25	0.25	0.26	0.20	0.35	0.20	95.3	95.0	95.1	95.9	94.6	93.5	94.6
	2.94	100	0.55	0.54	0.55	0.56	1.07	2.07	0.37	94.2	94.1	94.7	93.8	95.4	97.9	95.0
		200	0.36	0.36	0.36	0.37	0.74	1.42	0.26	94.9	94.1	94.8	95.5	95.4	96.5	94.4
		400	0.25	0.24	0.24	0.25	0.51	0.96	0.18	94.9	95.3	96.0	95.7	95.4	94.9	94.5
−2	0	100	0.62	0.62	0.61	0.64	0.43	0.76	0.45	92.4	93.6	92.6	93.8	94.6	93.5	92.7
		200	0.40	0.39	0.39	0.41	0.30	0.53	0.32	94.3	95.9	94.2	95.2	94.7	95.0	93.2
		400	0.27	0.27	0.27	0.28	0.21	0.37	0.22	95.4	95.2	94.1	93.7	94.6	94.1	93.5
	2.94	100	0.57	0.56	0.56	0.58	1.11	2.25	0.38	93.1	93.5	94.1	93.2	95.0	97.1	94.1
		200	0.37	0.37	0.37	0.38	0.78	1.52	0.27	93.9	95.6	93.9	94.6	94.6	95.9	94.3
		400	0.25	0.25	0.25	0.26	0.55	1.03	0.19	95.5	95.2	93.7	93.8	94.2	94.2	94.1

Note: The estimates are computed over 1000 repetitions of sample size $n = 100$, 200, and 400 for each combination of new infection rate μ, γ_0, and n.

detected across 152 isolates. More than 80% of initial infections contained multiple variants, displaying an average of 2.7 co-circulating variants. This analysis uses all the variants observed as the covariates for model building and classification. We assigned $\mu = -3$ in the reinfection probability model (4), the same as the northern Cambodia data.

Table 8 shows the classification result based on the EM algorithm and the two-stage method. We also report baseline and recurrence variants, classification probabilities, and classification results based on the posterior probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ and two-stage method $\hat{\xi}_i^{(1)}$. Two methods agree in every pair. The pair is classified as relapse when there is a high degree of overlap in dominant variants. For those pairs classified as reinfection, at least one non-sharing prevalent variant appeared in the recurrence sequencing. For example, in pair 30 → 30R, the variants F and S that appeared in the recurrence sequencing are unobserved in the baseline sequencing. Our algorithm classifies the recurrence as reinfection. Besides, the classifier using the recurrence sequencing information performs better than the classifier using only baseline sequencing. In pair 4 → 4R, for example, the classification probability π_{i2} is low, and the classifier would classify the infection as reinfection if one uses only baseline information. The posterior probability increases from 0.12 to 0.99 due to multiple overlapped variants between baseline and recurrence sequencing results.

Table 9 shows the estimation of regression coefficients by the EM algorithm. We report variants' prevalence, regression coefficient estimates, standard errors, and P-value. We use every variant observed and assess the significance using a

TABLE 6 Classification of recurrence pairs based on the EM algorithm and two-stage method for the northern Cambodia data

Recurrence pair	Baseline variants	$\pi_{i2}(x_i, \hat{\theta})$	Recurrence variants	$P(Y_i = 2 \mid \mathcal{O}_i, \hat{\theta})$	EM class	Two-stage class
10 → 10 R	A	0.17	*A*	0.89	Relapse	Relapse
31 → 31 R	A C E	0.79		1.00	Relapse	Relapse
36 → 36 R	A B C D E F G H	0.27	B C H	0.92	Relapse	Relapse
68 → 68 R	A *C* E	0.79		1.00	Relapse	Relapse
80 → 80 R	*A* E F I	0.63	A B C D F G H I	0.84	Relapse	Reinfection
81 → 81 R	A *B*	0.49	A *B*	0.99	Relapse	Relapse
82 → 82 R	A D E	0.64	A B *D*	0.98	Relapse	Relapse
87 → 87 R	A B *C* I	0.70	*A* H I	0.99	Relapse	Relapse
89 → 89 R	A E G I	0.90	B	1.00	Relapse	Reinfection
96 → 96 R	*A* C E I	0.93	A D	1.00	Relapse	Relapse
112 → 112 R	A B C E H	0.95	*A* B C	1.00	Relapse	Relapse
118 → 118 R	*I*	0.00	B *C*	0.00	Reinfection	Reinfection
123 → 123 R	*A* C	0.13	*A* B	0.82	Relapse	Reinfection
125 → 125 R	*C*	0.00	A B C E	0.00	Reinfection	Reinfection
126 → 126 R	A B C D E F G H	0.27	B *H*	0.96	Relapse	Relapse
130 → 130 R	A C D E	0.56	A E	0.97	Relapse	Relapse
151 → 151 R	D F *I*	0.00	A *I*	0.00	Reinfection	Reinfection
152 → 152 R	A *B*	0.49	*A* B F H	0.97	Relapse	Reinfection
153 → 153 R	A *E* H	0.86	*C*	0.99	Relapse	Relapse
154 → 154 R	*A* G	0.10	D *F* G	0.81	Relapse	Reinfection
160 → 160 R	C *E* H	0.00	A D F	0.00	Reinfection	Reinfection
177 → 177 R	A *E* H	0.86	*B*	0.99	Relapse	Relapse
179 → 179 R	D F H	0.00	B	0.00	Reinfection	Reinfection

Note: Dominant variants with a frequency of more than 50% are presented in italic.

TABLE 7 Estimation of regression coefficients by the EM algorithm for the northern Cambodia data

Variants	Prevalence	$\hat{\beta}$	SE	*P*-value
A	0.590	10.86	18.68	0.561
B	0.269	1.55	1.02	0.129
C	0.410	−0.32	0.90	0.722
D	0.295	−1.10	1.03	0.284
E	0.346	3.26	1.19	0.006
F	0.231	−2.32	1.30	0.076
G	0.231	−0.63	1.19	0.600
H	0.192	0.14	1.05	0.891
I	0.154	1.19	1.34	0.376

TABLE 8 Classification based on the EM algorithm and two-stage method for the southern Cambodia data

Recurrence pair	Baseline variants	$\pi_{i2}(x_i, \hat{\theta})$	Recurrence variants	$P(Y_i = 2 \mid \mathcal{O}_i, \hat{\theta})$	EM class	Two-stage class
4 → 4 R	D *S*	0.12	D *S*	0.99	Relapse	Relapse
7 → 7 R	H	0.67	H	1.00	Relapse	Relapse
10 → 10 R	H *S*	0.75	S	1.00	Relapse	Relapse
11 → 11 R	C F *S*	0.26	C D *F* G *S*	0.03	Reinfection	Reinfection
16 → 16 R	C *H* S	0.64	C H	1.00	Relapse	Relapse
17 → 17 R	C *H*	0.55	H	1.00	Relapse	Relapse
20 → 20 R	D G H S	0.89	D G S	1.00	Relapse	Relapse
25 → 25 R	C E F *G*	0.46	C G	0.98	Relapse	Relapse
27 → 27 R	C D *G* S	0.27	G *S*	0.98	Relapse	Relapse
28 → 28 R	C E G H	0.82	C D *G*	0.70	Relapse	Relapse
29 → 29 R	F *S*	0.36	F *S*	1.00	Relapse	Relapse
30 → 30 R	G	0.43	F S	0.04	Reinfection	Reinfection
42 → 42 R	G H	0.91	G	1.00	Relapse	Relapse
44 → 44 R	C D F G *S*	0.48	C D E F G *M*	0.83	Relapse	Relapse
47 → 47 R	D G	0.30	D E *S*	0.06	Reinfection	Reinfection
49 → 49 R	C D F H *S*	0.72	C D *H* S	1.00	Relapse	Relapse
50 → 50 R	H	0.67	H	1.00	Relapse	Relapse
51 → 51 R	D H *S*	0.63	H S	1.00	Relapse	Relapse
52 → 52 R	F *S*	0.36	S	0.99	Relapse	Relapse
54 → 54 R	F	0.29	G *S*	0.02	Reinfection	Reinfection
56 → 56 R	C D *G*	0.20	C	0.82	Relapse	Relapse
57 → 57 R	E G *S*	0.45	D G *S*	0.78	Relapse	Relapse
59 → 59 R	E G *M N*	1.00	E G *M N*	1.00	Relapse	Relapse
62 → 62 R	D F G *S*	0.61	D G *S*	1.00	Relapse	Relapse
63 → 63 R	D H S	0.63	D H S	1.00	Relapse	Relapse
66 → 66 R	C *F* G S	0.62	C F	0.98	Relapse	Relapse
68 → 68 R	*E* G	0.36	C D G H *S*	0.00	Reinfection	Reinfection
70 → 70 R	C *D* G S	0.27	C *S*	0.85	Relapse	Relapse
74 → 74 R	H	0.67	E F N S	0.00	Reinfection	Reinfection
75 → 75 R	E F H	0.80	E	1.00	Relapse	Relapse
77 → 77 R	C D	0.05	C D	0.91	Relapse	Relapse
78 → 78 R	C *H*	0.55	C *H*	1.00	Relapse	Relapse
80 → 80 R	F	0.29	F	1.00	Relapse	Relapse
83 → 83 R	C *F*	0.19	F	0.99	Relapse	Relapse
84 → 84 R	S	0.19	C D *S*	0.01	Reinfection	Reinfection
85 → 85 R	C *S*	0.12	C *S*	0.97	Relapse	Relapse
90 → 90 R	G	0.43	G	1.00	Relapse	Relapse
95 → 95 R	D *E F* G R	0.80	D *F* R	1.00	Relapse	Relapse
100 → 100 R	C D G *R*	0.55	C *G*	0.97	Relapse	Relapse
101 → 101 R	C E G *K* N	0.56	C E *N*	1.00	Relapse	Relapse
104 → 104 R	H	0.67	H	1.00	Relapse	Relapse
105 → 105 R	H	0.67	H	1.00	Relapse	Relapse
107 → 107 R	B C F G	0.18	C F	0.89	Relapse	Relapse
109 → 109 R	C F G *H* S	0.95	C D G *H*	0.97	Relapse	Relapse

Note: Dominant variants are presented in italic.

TABLE 9 Estimation of regression coefficients by the EM algorithm for the southern Cambodia data

Variants	Prevalence	$\hat{\beta}$	SE	*P*-value
A	0.007	−18.05	8619.58	0.998
B	0.033	−1.63	1.44	0.257
C	0.582	−0.51	0.56	0.36
D	0.399	−0.55	0.61	0.365
E	0.137	−0.27	1.15	0.817
F	0.294	0.92	0.60	0.127
G	0.320	1.55	0.61	0.011
H	0.229	2.56	0.69	<0.001
K	0.026	0.16	1.61	0.918
L	0.013	−12.84	1499.86	0.993
M	0.020	16.88	6182.29	0.998
N	0.039	1.14	1.80	0.527
Q	0.007	−14.64	2239.05	0.995
R	0.039	1.56	1.31	0.234
S	0.444	0.36	0.60	0.543
T	0.007	−16.62	8318.23	0.998

Wald-type test. As one can see, variants C, S, D, G, F, H, and E have a relatively high prevalence, but only variants G and H are statistically significant. Both variants have a positive association with the relapse. When variant G or H is presented in the baseline sequencing, the odds ratio of relapse is $\exp(1.55) = 4.7$ and $\exp(2.56) = 12.9$, respectively. To estimate the parameters related to transition probabilities, we have $\hat{\gamma}_1 = 0.64$, indicating that the presence of dominant variants at the baseline is positively associated with the transition probability of the variant in relapse. Other transition parameter estimates include $\hat{\gamma}_0 = 0.59$ and $\hat{\gamma}_0^* = 4.65$, indicating that both present and absent variants in the baseline sequencing would likely remain the same in the relapse. The corresponding standard errors are 0.23 and 0.51, respectively, showing a statistical significance from an utterly random transition. Some rare variants have a large standard error in Table 9. We added a sensitivity analysis in the supplementary material excluding those rare variants. Both estimation and classification results are similar to those using all variants.

We assume the occurrence of the variants is independent. This assumption has been checked for the northern Cambodia data in Lin et al[20] using Fisher's exact tests. We used the same approach for the southern Cambodia data in any 2 of the 16 variants. The result shows that only 10 out of 120 pairs have *P*-value smaller than 0.05. After adjusting for multiple comparisons using the Benjamini-Hochberg procedure, only one *P*-value is smaller than 0.05. The independence assumption is not significantly violated in the southern Cambodia data. The reinfection parameter is assigned as $\mu = -3$ in both data analyses. We added a sensitivity analysis in the supplementary material using $\mu = -2$. The classification result remains robust.

6 | DISCUSSION

This article proposes a novel classification methodology that utilizes the EM algorithm and transition likelihoods to classify the disease outcome. We applied the method to classify the recurrent *P vivax* infections as either relapse or reinfection in two Cambodian malaria research datasets. Compared to the previous method proposed by Lin et al,[20] our method has higher accuracy, especially when the sample size is small. Both simulation studies and real data analysis support our classifier's feasibility and practical use. Additionally, we generalize our method to include a transition model for probabilities q_{ij} and q_{ij}^* with external covariates. In our case, the variant's reading frequency can be used to model the transition.

In summary, the EM algorithm is more advantageous because of the simultaneous optimization of the joint likelihood functions, resulting in a more accurate classifier than the previous one using a two-stage method. The benefit is especially significant when the sample size is small. When the sample size is large, the two-stage method becomes more competitive and has a similar classification performance. In addition, the estimator derived from the EM algorithm has better finite-sample properties with a smaller empirical standard error and better variance estimation than the bootstrapping procedure, leading to better coverage of the true parameter. The better performance also occurs when the sample size is small. When the sample size is large, the two-stage estimator with bootstrapped variance estimation improves but may still be conservative and have a larger coverage probability than the nominal level. Since the two-stage method enjoys the advantage of a quicker convergence, one can use the estimates of the two-stage method as the starting values for the EM algorithm.

There are several avenues for further research in the applications of EM algorithms to infectious diseases. Our current analysis considers only recurrence indicators without the time domain involved, unlike the critical feature of DPM tracking disease severity over time. Our recent publication shows that the causes of the recurrent infection can be seen as competing risks, for which one observes the event occurrence of either relapse or reinfection.[30] Our simulations show that the EM algorithm with posterior classification probability can be a better classifier than the two-stage method when applied to the competing risk data with the missing cause of failure. Extending our algorithm to deal with competing risk data has the potential. However, our estimator is sensitive to selecting background disease occurrence rate (reinfection rate in the *P vivax* infection). If the background occurrence rate is misidentified, the maximum likelihood estimator of the coefficients may not be consistent, as indicated in Lin et al.[20] The classification probability $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ can also be biased. Multiple analyses under different values of background occurrence rates shall be implemented to explore the classification result's robustness. Taylor et al[14] used strongly informative priors for recrudescence and reinfection rates in a Bayesian framework, which can be considered an alternative approach to avoid the identification problem.

The EM algorithm performs well when applied to the low dimension data. Maximizing the Q function under a low-dimension scenario is straightforward and can be implemented by an optimization algorithm such as the Nelder-Mead method. However, when the dimension is high, for example, $J > n$, or many covariates are included in the transition model for q and q^*, the Q function's direct maximization is difficult or impossible. One may maximize the Q function with L_1 or L_2 penalty that shrinks the parameters in θ. Using our classifier, it is unclear how the EM algorithm would perform under the shrinkage method in the high-dimension scenario. We will leave it for future research.

Other extensions, such as spatial heterogeneity, are worthy of pursuing. For example, the time interval from primary infection to relapse in the *P vivax* infection may depend on the geographic location. The relapse often takes longer for people in temperate climes than tropical climes.[31,32] The difference between regions provides external information to extend the classification capacity. Also, in the parent study of the northern Cambodia data set, six participants suffered a second recurrent infection, and one participant suffered a third recurrent infection. We excluded those recurrent infections from our current analysis to simplify the interpretation. One shall extend the current methodology to recurrent events data with multiple event types using duration information between recurrent infections. Including the recurrent information shall improve the transition modeling.

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and three reviewers for their valuable comments. The National Institutes of Health partially support this work through grant award numbers UL1TR002489, R01AG073259, and K08AI110651.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Huijun Jiang https://orcid.org/0000-0002-7918-8805
Quefeng Li https://orcid.org/0000-0003-0707-2763
Jessica T. Lin https://orcid.org/0000-0002-4516-723X

Feng-Chang Lin https://orcid.org/0000-0002-2638-1775

REFERENCES

1. CDC. Self-study course: principles of epidemiology in public health practice. *Morb Mortal Wkly Rep*. 2006;55(42):1154.
2. Holford NHG, Sheiner LB. Understanding the dose-effect relationship: clinical application of pharmacokinetic-pharmacodynamic models. *Clin Pharm*. 1981;6(6):429-453.
3. Vogelstein B, Fearon ER, Hamilton SR, et al. Genetic alterations during colorectal-tumor development. *New Engl J Med*. 1988;319(9):525-532.
4. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schäffer AA. Inferring tree models for oncogenesis from comparative genome hybridization data. *J Comput Biol*. 1999;6(1):37-51.
5. Desper R, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH, Schäffer AA. Distance-based reconstruction of tree models for oncogenesis. *J Comput Biol*. 2000;7(6):789-803.
6. Beerenwinkel N, Rahnenführer J, Däumer M, et al. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*. 2005;12(6):584-598.
7. Gerstung M, Baudis M, Moch H, Beerenwinkel N. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*. 2009;25(21):2809-2815.
8. Tofigh A. *Using Trees to Capture Reticulate Evolution: Lateral Gene Transfers and Cancer Progression*. PhD thesis. KTH School of Computer Science and Communications; 2009; Stockholm, Sweden.
9. Bureau A, Shiboski S, Hughes JP. Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Stat Med*. 2003;22(3):441-462.
10. Hjelm M, Höglund M, Lagergren J. New probabilistic network models and algorithms for oncogenesis. *J Comput Biol*. 2006;13(4):853-865.
11. Robinson LJ, Wampfler R, Betuela I, et al. Strategies for understanding and reducing the plasmodium vivax and plasmodium ovale hypnozoite reservoir in papua new guinean children: a randomised placebo-controlled trial and mathematical model. *PLoS Med*. 2015;12(10):e1001891.
12. Lacerda MVG, Llanos-Cuentas A, Krudsood S, et al. Single-Dose tafenoquine to prevent relapse of plasmodium vivaxMalaria. *N Engl J Med*. 2019;380(3):215-228.
13. Llanos-Cuentas A, Lacerda MVG, Hien TT, et al. Tafenoquine versus primaquine to prevent relapse of plasmodium vivaxMalaria. *N Engl J Med*. 2019;380(3):229-241.
14. Taylor AR, Watson JA, Chu CS, et al. Resolving the cause of recurrent Plasmodium vivax malaria probabilistically. *Nat Commun*. 2019;10(1):5595.
15. Plucinski MM, Morton L, Bushman M, Dimbu PR, Udhayakumar V. Robust algorithm for systematic classification of malaria late treatment failures as recrudescence or reinfection using microsatellite genotyping. *Antimicrob Agents Chemother*. 2015;59:6096-6100.
16. Jones S, Plucinski M, Kay K, Hodel EM, Hastings IM. A computer modelling approach to evaluate the accuracy of microsatellite markers for classification of recurrent infections during routine monitoring of antimalarial drug efficacy. *Antimicrob Agents Chemother*. 2020;64:e01517-e01519.
17. Nyachieo A, van Overmeir C, Laurent T, Dujardin JC, D'Alessandro U. Plasmodium falciparum genotyping by microsatellites as a method to distinguish between recrudescent and new infections. *Am J Trop Med Hygiene*. 2005;73(1):210.
18. Kwiek JJ, Alker AP, Wenink EC, Chaponda M, Kalilani LV, Meshnick SR. Estimating true antimalarial efficacy by heteroduplex tracking assay in patients with complex plasmodium falciparum infections. *Antimicrob Agents Chemother*. 2007;51(2):521-527.
19. Lin JT, Hathaway NJ, Saunders DL, et al. Using amplicon deep sequencing to detect genetic signatures of plasmodium vivax relapse. *J Infect Dis*. 2015;212(6):999-1008.
20. Lin FC, Li Q, Lin JT. Relapse or reinfection: classification of malaria infection using transition likelihoods. *Biometrics*. 2020;76(4):1351-1363.
21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Methodol*. 1977;39(1):1-38.
22. Driss A, Hibbert JM, Wilson NO, Iqbal SA, Adamkiewicz TV, Stiles JK. Genetic polymorphisms linked to susceptibility to malaria. *Malar J*. 2011;10(1):271.
23. Bango ZA, Tawe L, Muthoga CW, Paganotti GM. Past and current biological factors affecting malaria in the low transmission setting of Botswana: a review. *Infect Genet Evol*. 2020;85:104458.
24. Louis TA. Finding the observed information matrix when using the EM algorithm. *J R Stat Soc B Methodol*. 1982;44(2):226-233.
25. Lon C, Manning JE, Vanachayangkul P, et al. Efficacy of two versus three-day regimens of dihydroartemisinin-piperaquine for uncomplicated malaria in military personnel in Northern Cambodia: an open-label randomized trial. *PLoS One*. 2014;9(3):e93138.
26. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing patterns of selection and geospatial genetic diversity within two leading plasmodium vivax candidate vaccine antigens. *PLoS Negl Trop Dis*. 2014;8(4):e2796.
27. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucl Acids Res*. 2018;46(4):e21.

28. Rogers WO, Sem R, Tero T, et al. Failure of artesunate-mefloquine combination therapy for uncomplicated Plasmodium falciparum malaria in southern Cambodia. *Malar J*. 2009;8(1):10.
29. Givens MB, Lin JT, Lon C, et al. Development of a capillary electrophoresis-based heteroduplex tracking assay to measure in-host genetic diversity of initial and recurrent plasmodium vivax infections in Cambodia. *J Clin Microbiol*. 2014;52(1):298-301.
30. Liu Y, Lin FC, Lin JT, Li Q. Classification of unknown cause of failure in competing risks: an application to recurrence of *Plasmodium vivax* malaria infection. *J Data Sci*. 2022;20:51-78.
31. White NJ. Determinants of relapse periodicity in plasmodium vivax malaria. *Malar J*. 2011;10(1):297.
32. Battle KE, Karhunen MS, Bhatt S, et al. Geographical variation in Plasmodium vivax relapse. *Malar J*. 2014;13(1):144.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Jiang H, Li Q, Lin JT, Lin F-C. Classification of disease recurrence using transition likelihoods with expectation-maximization algorithm. *Statistics in Medicine*. 2022;41(23):4697-4715. doi: 10.1002/sim.9534