Relapse or reinfection: Classification of malaria infection using transition likelihoods

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

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

[2]Institute of Global Health and Infectious Diseases, University of North Carolina, Chapel Hill, North Carolina

Correspondence
Feng-Chang Lin, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599.
Email: flin@bios.unc.edu

Funding information
National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR002489; National Institute of Allergy and Infectious Diseases (NIAID), Grant/Award Number: K08AI110651

Abstract

In patients with *Plasmodium vivax* malaria treated with effective blood-stage therapy, the recurrent illness may occur due to relapse from latent liver-stage infection or reinfection from a new mosquito bite. Classification of the recurrent infection as either relapse or reinfection is critical when evaluating the efficacy of an anti-relapse treatment. Although one can use whether a shared genetic variant exists between baseline and recurrence genotypes to classify the outcome, little has been suggested to use both sharing and nonsharing variants to improve the classification accuracy. In this paper, we develop a novel classification criterion that utilizes transition likelihoods to distinguish relapse from reinfection. When tested in extensive simulation experiments with known outcomes, our classifier has superior operating characteristics. A real data set from 78 Cambodian *P. vivax* malaria patients was analyzed to demonstrate the practical use of our proposed method.

KEYWORDS

allelic variant, amplicon deep sequencing, malaria relapse, penalized maximum likelihood estimation, *Plasmodium vivax*

1 | INTRODUCTION

The classification of infections from more than one potential cause is critical in malaria research. Taking *Plasmodium falciparum*, for example, the most prevalent malaria species in Sub-Saharan Africa, may recur due to relapse from treatment failure or due to reinfection from new mosquito bites. The true anti-malarial treatment efficacy cannot be determined without knowing whether the recurrent infection is due to treatment failure or new infection in an area of high malaria transmission (Kwiek *et al.*, 2007; Daniels *et al.*, 2008; Juliano *et al.*, 2010). *Plasmodium vivax*, the leading cause of malaria outside Africa, may similarly recur due to treatment failure or reinfection. However, in many endemic areas such as Southeast Asian and Oceania, it often recurs due to relapse of hypnozoites reactivating from the liver, as most anti-malarials are not active against these latent liver stages of *P. vivax* (Lin *et al.*, 2015; Beck *et al.*, 2016; Pearson *et al.*, 2016). Indeed, without knowing the cause of recurrent infection, determining treatment efficacy, relapse rate, and disease epidemiology is challenging.

Given the high degree of genetic diversity and polyclonal nature of *P. vivax* infections in many parts of the world, where many clones (genetically distinct strains) exist within a human host, a targeted amplicon deep sequencing approach provides an opportunity for a higher precision of classification (Lin *et al.*, 2015). As part of a malaria cohort study conducted from 2010 to 2011 (Lon *et al.*, 2014), patients in Cambodia found to have *P. vivax* were followed after treatment with a highly efficacious artemisinin-based combination therapy, with blood samples collected for deep sequencing at baseline. Of 78 infected patients followed, 23 individuals developed recurrent infections. Their blood sample was collected at the time of recurrence for another round of sequencing. It was hypothesized that through genotyping of the initial and recurrent parasite isolates, one may be able to distinguish relapse

from reinfection based on variant overlap between the two sequencing results within individuals.

Naively, one may classify the recurrent infection as relapse if any variant in the recurrent infection is shared with the initial infection (Nyachieo *et al.*, 2005; Kobbe *et al.*, 2006). However, without considering the prevalence of the variant, false positive misclassification likely occurs if some variants are frequently observed in the population (Juliano *et al.*, 2009). Kwiek *et al.* (2007) treated the recurrent infection as indeterminate if the initial and recurrent infections shared only one variant with a prevalence of more than 10%. However, this approach is somewhat ad hoc because the 10% prevalence cutoff may not be generally applied to other areas, and sharing only one variant may be rare in regions of high transmission where the parasite population is diverse, and a high number of variants is routinely detected in an individual (Juliano *et al.*, 2010). Instead, Lin *et al.* (2015) calculated the reinfection probability as the product of all reinfection probabilities from all shared variants and classified the recurrence as reinfection if the probability is more than 10%. Specifically, they calculated the reinfection probability based on a binomial probability model (BPM) that equals to $\prod_j \{1-(1-y_j)^x\}$, where y_j is the prevalence of a shared variant and x is the number of variants observed in the recurrent infection. As one can see, the probability model considers only the possibility of shared variants occurred in the recurrence. A nonshared variant may also occur at random in the recurrent infection, regardless whether the recurrence is relapse or reinfection. This is likely due to reactivation of latent parasites acquired from other, historical infections preceding those captured by genotyping (Chen *et al.*, 2007; Imwong *et al.*, 2007).

The presence or absence of variants in the initial and recurrence sequencings can naturally be described by a transition model. However, the estimation of transition probabilities is complicated by an unknown mixture of two models, one from relapse and one from reinfection. Here in this paper, we propose an estimation procedure that can estimate the transition probabilities under unknown causes of infections. The method is first established on a statistical model that can describe the probability of relapse in the recurrent infection. Then, through comparison of two transition likelihoods, our novel classification criterion utilizing the transition information can significantly improve a classifier that uses only initial sequencing information.

The rest of the paper is organized as follows. In Section 2, we develop a statistical model for the probability of observing a recurrent infection in the follow-up period, which sums over probabilities of relapse and reinfection. A likelihood-based estimation method is utilized, with a computing solution for high-dimensional data when the number of allelic variants exceeds the number of subjects. Our novel classification criterion is discussed in Section 3. Simulation studies in Section 4 for both low- and high-dimensionality scenarios show the consistency and high accuracy of our classifier. A comparison to the existing BPM method (Lin *et al.*, 2015) shows the superiority of our approach. We apply our method to the *P. vivax* infection data and present part of the classification results in Section 5. Assumptions and possible generalizations of our approach are discussed in Section 6.

2 | STATISTICAL MODEL AND ESTIMATION

2.1 | Notation

For subject i, let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})'$ and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iJ})'$ denote a J-dimensional vector of sequencing outcomes in the initial and recurrent infections, respectively. Let Y_i be the binary indicators with $Y_i = 1$ indicating the recurrence and 0 otherwise. In this study, we aim to classify the recurrent infection, $Y_i = 1$, into two latent classes, namely, relapse $R_i = 1$ or reinfection $N_i = 1$, assuming that two types of infections cannot occur simultaneously. We also assume that a third possible class, treatment failure, is unlikely in the setting of highly efficacious therapy. Note that the sequencing outcomes $\boldsymbol{Z}_i$ in the recurrent infection can only be observed when $Y_i = 1$, and can be different from $\boldsymbol{X}_i$ even when the recurrent infection is relapse. If a subject does not have a recurrent infection, that is, $Y_i = 0$, the information on $\boldsymbol{Z}_i$ is not available. Through the paper, the number of subjects from the baseline with initial sequencing is denoted by n, and the number of subjects who have recurrent infections with follow-up sequencing is denoted by $m = \sum_{i=1}^{n} Y_i$.

2.2 | Statistical model

Suppose that $P(X_{ij} = 1) = p_j$ for $i = 1, \ldots, n$, where $X_{ij} = 1$ indicates the presence of variant j in the sequencing outcome of subject i, and $X_{ij} = 0$ otherwise. Given a realization of the initial sequencing outcome $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})'$, the indicator for relapse, denoted by R_i, is assumed to follow a logistic model

$$\log\left\{\frac{\pi_i(\theta)}{1-\pi_i(\theta)}\right\} = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{1}$$

where $\pi_i(\theta) = P(R_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$, $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}')'$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ with $\beta_j = 0$ indicating the jth variant is not associated with the relapse.

However, the relapse indicator R_i cannot be observed. What can be observed is the recurrence indicator Y_i, which equals 1 if either relapse or reinfection occurs. Assume that

the probability of acquiring an reinfection is constant and independent of the baseline variants $\boldsymbol{X}_i$, that is,

$$P(N_i = 1|R_i = 0, \boldsymbol{X}_i) = P(N_i = 1|R_i = 0) = \mu,$$

and that both infections cannot occur simultaneously, that is,

$$P(N_i = 1|R_i = 1, \boldsymbol{X}_i) = 0.$$

One can write

$$\begin{aligned} &P(Y_i = 1|\boldsymbol{X}_i) \\ &= P(N_i = 1, R_i = 0|\boldsymbol{X}_i) + P(N_i = 0, R_i = 1|\boldsymbol{X}_i) \\ &= P(N_i = 1|R_i = 0, \boldsymbol{X}_i)P(R_i = 0|\boldsymbol{X}_i) \\ &\quad + P(N_i = 0|R_i = 1, \boldsymbol{X}_i)P(R_i = 1|\boldsymbol{X}_i) \\ &= \mu\{1 - \pi_i(\boldsymbol{\theta})\} + \pi_i(\boldsymbol{\theta}), \end{aligned} \tag{2}$$

where $\pi_i(\boldsymbol{\theta}) = \exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta})/\{1 + \exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta})\}$ as defined in model (1).

Note that assuming constant infection rate is reasonable because subjects who live in the same area shall be bite by mosquitoes completely at random. The reinfection rate may depend on risk factors. If so, we may build a regression model relating the reinfection rate to those risk factors. Our approach still applies after such adjustment.

2.3 | Estimation method

For the binary outcome Y_i, one can estimate the unknown parameters minimizing the negative log-likelihood function. However, the reinfection probability μ and baseline relapse rate α may not be estimable simultaneously because both parameters are part of the baseline recurrences. To avoid such identifiability problem, we assume the reinfection rate is known or can be estimated via external information. Given μ, the parameter $\boldsymbol{\theta}$ can be estimated via minimizing the negative log-likelihood function

$$\ell(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}[y_i \log p_i(\boldsymbol{\theta}) + (1 - y_i)\log\{1 - p_i(\boldsymbol{\theta})\}],$$

where y_i is a realization of Y_i and $p_i(\boldsymbol{\theta}) = \mu\{1 - \pi_i(\boldsymbol{\theta})\} + \pi_i(\boldsymbol{\theta})$. Under regularity conditions for maximum likelihood estimators, one can show that $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ and $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a normal variable with mean 0 and variance that is the inverse of Fisher information matrix.

In our data where the number of variants is larger than the number of patients, we penalize the likelihood function with an L_1-penalty (Tibshirani, 1996) to enable variable selection and avoid ill-posed minimization problem when $J > n$. In particular, we solve the following optimization problem

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} -\ell(\boldsymbol{\theta}) + \lambda\sum_{j=1}^{J}|\beta_j|, \tag{3}$$

where λ is a tuning parameter whose optimal value will be determined by cross-validation. There are some other choices of penalty functions, such as elastic net penalty (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), or folded-concave penalty (Fan and Lv, 2011). From numerical studies, we found that the performance of our method is not sensitive to the choice of penalty functions. The main purpose of penalization is to regulate the optimization problem with high-dimensional covariates and select baseline variants that associate with recurrence.

2.4 | Computation

We develop a coordinate gradient descent algorithm (Friedman *et al.*, 2010) to solve the optimization problem (3). Let $\widetilde{\boldsymbol{\theta}} = (\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}')'$ be the current value of $\boldsymbol{\theta}$ and $\widetilde{\vartheta}_i = \widetilde{\alpha} + \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}$. Let $f(\vartheta_i) = y_i \log p(\vartheta_i) + (1 - y_i)\log\{1 - p(\vartheta_i)\}$ with $\vartheta_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta}$, and let $f'(\vartheta_i)$ and $f''(\vartheta_i)$ denote the first and second derivatives of the function f with respect to ϑ_i, respectively. A local quadratic approximation to $-\ell(\boldsymbol{\theta})$ can be written as

$$\begin{aligned} &\ell_Q(\boldsymbol{\theta}; \widetilde{\boldsymbol{\theta}}) \\ &= n^{-1}\sum_{i=1}^{n}\left\{-\frac{1}{2}f''(\widetilde{\vartheta}_i)(\vartheta_i - \widetilde{\vartheta}_i)^2 - f'(\widetilde{\vartheta}_i)(\vartheta_i - \widetilde{\vartheta}_i)\right\} + c_1(\widetilde{\boldsymbol{\theta}}) \\ &= (2n)^{-1}\sum_{i=1}^{n} -f''(\widetilde{\vartheta}_i)\left\{\vartheta_i - \widetilde{\vartheta}_i + \frac{f'(\widetilde{\vartheta}_i)}{f''(\widetilde{\vartheta}_i)}\right\}^2 + c_2(\widetilde{\boldsymbol{\theta}}) \\ &= (2n)^{-1}\sum_{i=1}^{n}\widetilde{\omega}_i(\widetilde{\vartheta}_i^* - \alpha - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + c_2(\widetilde{\boldsymbol{\theta}}), \end{aligned}$$

where $\widetilde{\vartheta}_i^* = \widetilde{\vartheta}_i - f'(\widetilde{\vartheta}_i)/f''(\widetilde{\vartheta}_i)$, $\widetilde{\omega}_i = -f''(\widetilde{\vartheta}_i)$, and $c_1(\widetilde{\boldsymbol{\theta}})$ and $c_2(\widetilde{\boldsymbol{\theta}})$ are functions depending only on $\widetilde{\boldsymbol{\theta}}$. We then minimize $\ell_Q(\boldsymbol{\theta}; \widetilde{\boldsymbol{\theta}}) + \lambda\sum_{j=1}^{J}|\beta_j|$, which becomes a regularized weighted least squares problem:

$$\widetilde{\boldsymbol{\theta}}^{\text{new}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\,(2n)^{-1}\sum_{i=1}^{n}\widetilde{\omega}_i(\widetilde{\vartheta}_i^* - \alpha - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{J}|\beta_j|. \tag{4}$$

Such a problem can be solved by a standard coordinate gradient descent algorithm (Friedman *et al.*, 2010), which is implemented by R package `glmnet`.

The remaining tasks are to derive $f'(\vartheta_i)$ and $f''(\vartheta_i)$. By the definition of $p(\vartheta_i)$, one has $p'(\vartheta_i) = (1-\mu)\pi_i(\boldsymbol{\theta})\{1-\pi_i(\boldsymbol{\theta})\}$, and $p''(\vartheta_i) = (1-\mu)\pi_i(\boldsymbol{\theta})\{1-\pi_i(\boldsymbol{\theta})\}\{1-2\pi_i(\boldsymbol{\theta})\}$. Then, one can write $f'(\vartheta_i)$ as

$$\begin{aligned} f'(\vartheta_i) &= \frac{y_i}{p(\vartheta_i)}p'(\vartheta_i) - \frac{1-y_i}{1-p(\vartheta_i)}p'(\vartheta_i) \\ &= \frac{p'(\vartheta_i)}{p(\vartheta_i)\{1-p(\vartheta_i)\}}\{y_i - p(\vartheta_i)\}. \end{aligned} \tag{5}$$

Since

$$\begin{aligned} \log f'(\vartheta_i) &= \log p'(\vartheta_i) + \log\{y_i - p(\vartheta_i)\} \\ &\quad - \log p(\vartheta_i) - \log\{1-p(\vartheta_i)\}, \end{aligned}$$

taking derivatives on both sides gives

$$\begin{aligned} \frac{f''(\vartheta_i)}{f'(\vartheta_i)} &= \frac{\partial}{\partial\vartheta_i}\log f'(\vartheta_i) = \frac{p''(\vartheta_i)}{p'(\vartheta_i)} \\ &\quad - \left\{\frac{1}{y_i - p(\vartheta_i)} + \frac{1}{p(\vartheta_i)} - \frac{1}{1-p(\vartheta_i)}\right\}p'(\vartheta_i). \end{aligned} \tag{6}$$

Straightforwardly, the product of (5) and (6) gives $f''(\vartheta_i)$. We summarize the algorithm as follows:

Step 1: Initialize $\boldsymbol{\theta}$ at $\widetilde{\boldsymbol{\theta}} = (\widetilde{\alpha}, \widetilde{\boldsymbol{\beta}}')'$.

Step 2: Solve

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}^{\text{new}} &= \underset{\theta}{\text{argmin}}\,(2n)^{-1}\sum_{i:\widetilde{\omega}_i>0}\widetilde{\omega}_i(\widetilde{\vartheta}_i^* - \alpha - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \\ &\quad + \lambda\sum_{j=1}^{J}|\beta_j|, \end{aligned}$$

where $\widetilde{\vartheta}_i^* = \widetilde{\vartheta}_i - f'(\widetilde{\vartheta}_i)/f''(\widetilde{\vartheta}_i)$, $\widetilde{\vartheta}_i = \widetilde{\alpha} + \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}$, and $\widetilde{\omega}_i = -f''(\widetilde{\vartheta}_i)$.

Step 3: Update $\widetilde{\vartheta}_i$, $\widetilde{\vartheta}_i^*$, and $\widetilde{\omega}_i$ by letting $\widetilde{\vartheta}_i^{\text{new}} = \widetilde{\alpha}^{\text{new}} + \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}^{\text{new}}$, $\widetilde{\vartheta}_i^{*\text{new}} = \widetilde{\vartheta}_i^{\text{new}} - f'(\widetilde{\vartheta}_i^{\text{new}})/f''(\widetilde{\vartheta}_i^{\text{new}})$, and $\widetilde{\omega}_i^{\text{new}} = -f''(\widetilde{\vartheta}_i^{\text{new}})$.

Step 4: Iterate between steps 2 and 3 until convergence, that is, the L_2-norm $||\widetilde{\boldsymbol{\theta}}^{\text{new}} - \widetilde{\boldsymbol{\theta}}||_2 \le \epsilon$, where ϵ is a user-defined stopping threshold. We choose $\epsilon = 0.001$.

Remark that, when $\mu > 0$, the function $-f(\vartheta_i)$ is not a convex function. Therefore, solving our proposed target function (3) is a challenging nonconvex optimization problem. To ensure stable computation of the gradient descent algorithm, we drop negative weight $\widetilde{\omega}_i$ when solving the intermediate weighted least squares function (4) in Step 2 above. Similar to other nonconvex optimization problems, the gradient descent algorithm converges to a local minimum of the objective function. In the simulation studies, we find that such local minima admit good variable selection and classification performance; see Section 4.

3 | CLASSIFICATION

We aim to classify recurrent infection ($Y_i = 1$) to either relapse ($R_i = 1$) or reinfection ($N_i = 1$). Two classifiers are studied. The first one utilizes the initial sequencing information and logistic regression model (1) to calculate the initial probability estimation of the recurrence being relapse. The second one updates the initial probability estimation using transition likelihoods under relapse and reinfection. Through comparison between two transition likelihoods, the second classifier is anticipated to perform better than the first one because more information is used.

3.1 | Based on baseline information

Let ξ_i denote the probability of being relapse given that a recurrent infection has occurred. One can show that, based on the recurrence probability in formula (2),

$$\begin{aligned} \xi_i^{(0)} &= P(R_i = 1|Y_i = 1, \boldsymbol{X}_i) \\ &= P(N_i = 0|R_i = 1, \boldsymbol{X}_i)\frac{P(R_i = 1|\boldsymbol{X}_i)}{P(Y_i = 1|\boldsymbol{X}_i)} \\ &= \frac{\pi_i(\boldsymbol{\theta})}{\mu\{1-\pi_i(\boldsymbol{\theta})\} + \pi_i(\boldsymbol{\theta})}, \end{aligned}$$

which can be estimated by

$$\widehat{\xi}_i^{(0)} = \frac{\pi_i(\widehat{\boldsymbol{\theta}})}{\mu\{1-\pi_i(\widehat{\boldsymbol{\theta}})\} + \pi_i(\widehat{\boldsymbol{\theta}})}.$$

This estimator gives a possible classification criterion via ranking $\widehat{\xi}_i^{(0)}$. Acknowledging the interpretation of probability, one may claim the recurrent case is $(100 \times \widehat{\xi}_i^{(0)})$-percent likely to be relapse. However, one may ask for a clear cut to identify the relapse. Barring this in mind, one can classify a recurrent infection to be relapse if $\widehat{\xi}_i^{(0)} > 0.5$, which means $\pi_i(\widehat{\boldsymbol{\theta}}) > \mu\{1-\pi_i(\widehat{\boldsymbol{\theta}})\}$ or equivalently, $P(N_i = 0, R_i = 1|Y_i = 1, \boldsymbol{X}_i) > P(N_i = 1, R_i = 0|Y_i = 1, \boldsymbol{X}_i)$. The cutoff could be chosen to optimize the operating characteristics if the true infection type is available. Without the gold standard in this study, we simply use 0.5 as the cutoff to choose the winner.

3.2 | Updated by transition likelihoods

The variant present or absent in the baseline sequencing may not be present or absent again in the follow-up

sequencing. Recall that $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iJ})'$ is a random variable for the recurrence sequencing outcomes. Assuming the recurrent infection is a relapse, one can write Z_{ij} as

$$Z_{ij} = X_{ij}\delta_{ij} + (1 - X_{ij})(1 - \delta^*_{ij}),$$

where δ_{ij} and δ^*_{ij} are two binary indicators that represent repeated presence and absence of variant j in the recurrence sequencing, with probability $q_j = P(\delta_{ij} = 1)$ and $q^*_j = P(\delta^*_{ij} = 1)$, respectively. Specifically, we assume that variant j has probability $q_j = P(Z_{ij} = 1|X_{ij} = 1, R_i = 1)$ to be observed again in the recurrence sequencing if the variant is observed in the initial sequencing, while the variant has probability $q^*_j = P(Z_{ij} = 0|X_{ij} = 0, R_i = 1)$ to remain unobserved in the recurrence sequencing if the variant is absent at the baseline. This mechanism can be considered as a transition model from the baseline sequencing to the follow-up sequencing outcomes, where q_j and q^*_j are transition probabilities in a two-state transition model. If the recurrence is indeed a new infection, we assume $\boldsymbol{Z}_i$ is independent of $\boldsymbol{X}_i$, and follows the same distribution as $\boldsymbol{X}_i$.

When $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are observed, parameters p_j, q_j, and q^*_j, $j = 1, \ldots, J$, can be estimated as follows. The probability p_j can be consistently estimated by the sample mean $\widehat{p}_j = n^{-1}\sum_{i=1}^n x_{ij}$, where x_{ij} is a realization of X_{ij}. Different from the baseline variants, the distribution of Z_{ij} is a mixture of two distributions, depending on whether the recurrent case is relapse or reinfection. Assuming the variants are mutually independent, we have

$$\begin{aligned} &P(Z_{ij} = 1|R_i = 1, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\quad = x_{ij}q_j + (1 - x_{ij})(1 - q^*_j) = \eta(x_{ij}), \end{aligned}$$

$$\begin{aligned} &P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = 1, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\quad = \prod_{j=1}^J \eta(x_{ij})^{z_{ij}}\{1 - \eta(x_{ij})\}^{1-z_{ij}}, \end{aligned}$$

and

$$P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = 0, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) = \prod_{j=1}^J p_j^{z_{ij}}(1 - p_j)^{1-z_{ij}}.$$

Let $\phi^R_i(q, q^*) = P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = 1, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)$, where $q = (q_1, \ldots, q_J)'$ and $q^* = (q^*_1, \ldots, q^*_J)'$, and let $\phi^N_i(p) = P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = 0, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)$, where $p = (p_1, \ldots, p_J)'$. The mixture distribution of $\boldsymbol{Z}_i$ can be written as

$$\begin{aligned} &P(\boldsymbol{Z}_i = \boldsymbol{z}_i|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\quad = \sum_{r=0}^1 P(\boldsymbol{Z}_i = \boldsymbol{z}_i, R_i = r|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\quad = \sum_{r=0}^1 P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = r, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\qquad \times P(R_i = r|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ &\quad = \phi^N_i(p)(1 - \xi^{(0)}_i) + \phi^R_i(q, q^*)\xi^{(0)}_i. \end{aligned}$$

To obtain the maximum likelihood estimators for q and q^*, we maximize the profiled log-likelihood function

$$\ell(q, q^*) = \sum_{i=1}^m \log\{\phi^N_i(\widehat{p})(1 - \widehat{\xi}^{(0)}_i) + \phi^R_i(q, q^*)\widehat{\xi}^{(0)}_i\}, \quad (7)$$

where $\phi^N_i(\widehat{p}) = \prod_{j=1}^J \widehat{p}_j^{z_{ij}}(1 - \widehat{p}_j)^{1-z_{ij}}$, and $\widehat{\xi}^{(0)}_i$ is the estimated probability of relapse based on the baseline sequencing information.

Based on the transition model for the follow-up sequencing outcomes, one can derive the probability of relapse given the follow-up sequencing realization $\boldsymbol{Z}_i = \boldsymbol{z}_i$. One can show that,

$$\begin{aligned} \xi^{(1)}_i &= P(R_i = 1|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{z}_i) \\ &= \frac{\begin{array}{c} P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = 1, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ \times P(R_i = 1|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \end{array}}{\begin{array}{c} \sum_{r=0}^1 P(\boldsymbol{Z}_i = \boldsymbol{z}_i|R_i = r, Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \\ \times P(R_i = r|Y_i = 1, \boldsymbol{X}_i = \boldsymbol{x}_i) \end{array}} \\ &= \frac{\phi^R_i(q, q^*)\xi^{(0)}_i}{\phi^N_i(p)(1 - \xi^{(0)}_i) + \phi^R_i(q, q^*)\xi^{(0)}_i}, \end{aligned}$$

which can be consistently estimated by

$$\widehat{\xi}^{(1)}_i = \frac{\phi^R_i(\widehat{q}, \widehat{q}^*)\widehat{\xi}^{(0)}_i}{\phi^N_i(\widehat{p})(1 - \widehat{\xi}^{(0)}_i) + \phi^R_i(\widehat{q}, \widehat{q}^*)\widehat{\xi}^{(0)}_i},$$

where

$$\phi^R_i(\widehat{q}, \widehat{q}^*) = \prod_{j=1}^J \widehat{\eta}(x_{ij})^{z_{ij}}\{1 - \widehat{\eta}(x_{ij})\}^{1-z_{ij}},$$

and

$$\widehat{\eta}(x_{ij}) = x_{ij}\widehat{q}_j + (1 - x_{ij})(1 - \widehat{q}^*_j),$$

where $\widehat{q}_j$ and $\widehat{q}^*_j$ are maximum likelihood estimators solving (7).

The estimator provides another classifier as one may claim the recurrent case is $(100 \times \widehat{\xi}^{(1)}_i)$-percent likely to be relapse and classify the recurrence as relapse if $\widehat{\xi}^{(1)}_i > 0.5$. In fact, $\widehat{\xi}^{(1)}_i$ can be considered as the probability that updates $\widehat{\xi}^{(0)}_i$ by a ratio of two transition likelihoods $\phi^N_i(p)$ and $\phi^R_i(q, q^*)$. More

specifically, the estimated odds of relapse given the follow-up information can be written as

$$\widehat{\text{Odds}}_i^{(1)} = \frac{\widehat{\xi}_i^{(1)}}{1-\widehat{\xi}_i^{(1)}} = \frac{\phi_i^R(\widehat{q},\widehat{q}^*)\widehat{\xi}_i^{(0)}}{\phi_i^N(\widehat{p})(1-\widehat{\xi}_i^{(0)})} = \frac{\phi_i^R(\widehat{q},\widehat{q}^*)}{\phi_i^N(\widehat{p})}\widehat{\text{Odds}}_i^{(0)},$$

which updates the estimated odds from the baseline information by multiplying the ratio of two transition likelihoods. If $\phi_i^R(\widehat{q},\widehat{q}^*) > \phi_i^N(\widehat{p})$, the realization of $\boldsymbol{Z}_i$ more likely came from relapse. Hence, the odds of the recurrent infection being relapse would increase from the one that uses only baseline information.

Note that as $\boldsymbol{Z}_i$ is only available from m subjects who have recurrent infections, the parameters q and q^* cannot be solved by the likelihood function (7) when the combined dimensions of q and q^* is larger than the number of subjects m. To avoid this, we assume the transition probabilities are the same for each variant, that is, $q_1 = q_2 = \cdots = q_J$ and $q_1^* = q_2^* = \cdots = q_J^*$, such that there are only two scalar parameters q and q^* in (7). A possible generalization that relaxes this assumption is discussed in Section 6.

4 | SIMULATION EXPERIMENTS

In this section, we demonstrate our method via simulation experiments with various combinations of reinfection rate μ, sample size n, and number of variants J. First, we explore a low-dimension setting when there are only 10 variants in both sequencings. The baseline sequencing outcomes X_{ij}, $j = 1, \ldots, 10$, are assumed to follow a Bernoulli distribution with success probability $p_j = 0.5\exp\{-(j-1)/10\}$, which mimics the distribution of variants in our real data. Two transition probabilities, q_j and $\widetilde{q}_j$, are set to be 0.95. The probability of acquiring a new infection is set to be $\mu = 0.05, 0.12, 0.25$, from low to high reinfection rates. We explore two scenarios under which the association between the presence of the variant and relapse is different. In the first scenario, we assume that the relapse is associated with three most prevalent variants X_{i1}, X_{i2}, X_{i3}. In the second scenario, we assume the relapse is associated with three rarest variants X_{i8}, X_{i9}, X_{i10}. In each scenario, we set the intercept $\alpha = -2$ in the relapse model (1) and coefficients $\beta_j = 0.405$ if the variant j is associated with the relapse and $\beta_j = 0$ otherwise. The sample size is set to be $n = 100, 200, 400, 800$. We simulate 1000 repetitions for each combination of μ and n in each scenario. We report the bias of the coefficient estimates to demonstrate the consistency of our proposed estimator for the regression coefficients. We also report operating characteristics such as sensitivity (sens), specificity (spec), and overall accuracy (acc) of the classifiers $I(\widehat{\xi}^{(0)} > 0.5)$ and $I(\widehat{\xi}^{(1)} > 0.5)$. We also compare our method to the BPM used in Lin *et al.* (2015).

TABLE 1 Bias of regression coefficient estimation under scenario 1

μ	n	β_1	β_2	β_3	β_4	β_5
0.05	100	1.129	1.243	1.181	−0.142	−0.276
	200	0.114	0.044	0.093	−0.048	−0.081
	400	0.026	0.014	0.041	−0.003	0.012
	800	0.012	0.005	0.021	−0.004	0.007
0.12	100	3.899	3.774	3.837	−3.748	−0.519
	200	0.742	0.379	0.424	−0.050	−0.341
	400	0.028	0.030	0.045	−0.004	−0.010
	800	0.008	0.005	0.030	−0.014	0.001
0.25	100	5.816	4.482	4.579	−0.814	−2.132
	200	3.066	2.573	2.478	−0.985	−0.544
	400	0.271	0.092	0.079	0.120	−0.039
	800	0.014	0.028	0.029	0.001	0.007
μ	n	β_6	β_7	β_8	β_9	β_{10}
0.05	100	−0.756	−1.409	−1.295	−0.863	−0.754
	200	−0.089	−0.135	−0.197	−0.135	−0.153
	400	−0.022	−0.032	−0.028	−0.026	−0.043
	800	−0.003	−0.015	−0.012	−0.010	−0.023
0.12	100	−3.657	−1.947	−1.914	−1.718	−5.552
	200	−0.860	−0.973	−0.214	−2.913	−0.650
	400	−0.049	−0.053	0.004	−0.015	−0.062
	800	−0.018	−0.023	0.001	−0.001	−0.013
0.25	100	−4.738	−4.455	−3.278	−5.488	−4.870
	200	−1.155	−2.339	−3.635	−2.956	−1.537
	400	−0.143	−0.263	−0.225	−0.612	−0.210
	800	−0.016	−0.010	−0.072	−0.037	−0.020

Tables 1 and 2 show the simulation results under the scenarios 1 and 2, respectively. One can see that our estimator is consistent. When the sample size n increases, the bias converges toward 0. It is worth noting that our estimator performs equally well in those two scenarios when either common or rare variants are associated with the relapse. Table 3 shows the operating characteristics of three classifiers under different reinfection rates. One can see that using $I(\widehat{\xi}^{(0)} > 0.5)$ as the classifier can be overly aggressive under a low reinfection rate. Most of the recurrences are claimed as relapse and result in high sensitivity but low specificity, especially when the sample size is large. On the other hand, using $I(\widehat{\xi}^{(1)} > 0.5)$ as the classifier performs well, reaching a high degree of accuracy in both sensitivity and specificity. The reinfection rate is a significant factor for the classification accuracy of our classifiers. Under a high reinfection rate, the overall accuracy of the classifier $I(\widehat{\xi}^{(0)} > 0.5)$ is low. Correctly classifying relapse becomes more difficult for the classifier using $\widehat{\xi}^{(0)}$. The same problem occurs to $I(\widehat{\xi}^{(1)} > 0.5)$ when the sample size is small. However, when the sample size increases, the accuracy of $I(\widehat{\xi}^{(1)} > 0.5)$ increases to a satisfactory level, under either

TABLE 2 Bias of regression coefficient estimation under scenario 2

μ	n	β_1	β_2	β_3	β_4	β_5
0.05	100	0.656	0.367	0.177	−0.746	−0.924
	200	−0.100	−0.004	0.192	−0.178	−0.108
	400	0.010	−0.019	0.004	−0.015	−0.002
	800	0.009	−0.011	−0.004	−0.006	−0.006
0.12	100	−0.368	−1.593	−2.398	−1.302	−1.104
	200	−0.541	−0.682	−0.138	−1.069	−2.454
	400	−0.015	0.006	−0.050	−0.055	−0.002
	800	−0.012	−0.007	−0.004	−0.013	0.007
0.25	100	−1.367	−2.823	−1.996	−3.290	−1.771
	200	−1.325	−1.096	−1.609	−2.230	−3.028
	400	−0.467	−0.876	0.152	−1.062	−0.190
	800	0.012	−0.090	−0.002	−0.023	−0.019
μ	n	β_6	β_7	β_8	β_9	β_{10}
0.05	100	−1.965	−1.708	0.768	0.700	1.537
	200	−0.257	−0.220	0.072	−0.090	0.080
	400	−0.017	−0.033	0.008	−0.017	0.000
	800	−0.003	−0.010	0.006	−0.008	−0.001
0.12	100	−3.751	−6.018	1.511	0.675	−0.877
	200	−1.751	−1.306	0.509	0.586	0.267
	400	−0.074	−0.150	0.041	−0.029	−0.074
	800	−0.019	−0.008	0.006	0.001	0.001
0.25	100	−3.384	−3.217	2.677	1.502	1.927
	200	−4.561	−3.112	1.756	−0.204	1.783
	400	−1.325	−1.724	0.370	−0.042	0.158
	800	−0.293	−0.091	0.031	−0.078	−0.050

common or rare variants scenario. In comparison to the BPM method (Lin *et al.*, 2015), one can see that the BPM's performance remains unchanged under different reinfection rates. The method generally performs better than our classifier when the sample size is small. However, when the sample size is large or when the reinfection rate is low, our classifier performs much better than the BPM method. Note that the cutoff probability used in the BPM can be arbitrary and may depend on the prevalence of the shared variant, which is unknown in practice. It is not clear how to select the best cutoff for their method. We used 10% as suggested in Lin *et al.* (2015).

In addition, we performed a simulation study when the number of variants exceeds the sample size. We simulate baseline and follow-up sequencing outcomes from the same distribution as in the low-dimensional case. We consider two combinations of n and J for $(n, J) = (100, 200)$ and $(n, J) = (200, 400)$ and three reinfection rates for $\mu = 0.05, 0.12, 0.25$. We also consider two scenarios of how the variants associate with relapse. In scenario 3, the relapse is associated with five most prevalent variants through model (1), where $\alpha = -1$ and $\boldsymbol{\beta} = (0.2, 0.2, 0.2, 0.2, 0.2, 0, \ldots, 0)'$. In scenario 4, the relapse is associated with five relatively rare variants, where $\alpha = -1$ and $\boldsymbol{\beta} = (0, \ldots, 0, 0.2, 0.2, 0.2, 0.2, 0.2, 0, \ldots, 0)'$ with the 10 most prevalent variants not associated with relapse. Average sensitivity (sens), specificity (spec), and overall accuracy (acc) of the two classifiers are reported in Table 4 for each scenario. We also report average bias (Bias), which is defined by $(1/J)\sum_{j=1}^{J}|\widehat{\beta}_j - \beta_j|$, number of true positives (TP), and number of true negatives (TN) to evaluate the variable selection performance of our method.

Table 4 shows that our method still works well when J is much larger than n. The performance of the classifier $I(\widehat{\xi}_i^{(1)} > 0.5)$ is much better than that of the classifier $I(\widehat{\xi}_i^{(0)} > 0.5)$. When the reinfection rate is relatively high, both classifiers suffer lower accuracy under this more difficult situation. The performance of the classifier $I(\widehat{\xi}_i^{(1)} > 0.5)$, however, remains acceptable. Moreover, our method identifies most of the variants that are associated with the relapse, that is, its TP proportion is high, regardless of whether they are prevalent or rare, while controlling the selection of true negatives at a satisfactory level.

5 | REAL DATA ANALYSIS

Given the high degree of genetic diversity and polyclonal nature of *P. vivax* infections in Cambodia, many clones or strains exist within a human host. A targeted amplicon deep sequencing approach was chosen to genotype initial and recurrent isolates from *P. vivax* patients enrolled in a malaria cohort and treatment study conducted in northern Cambodia from 2010 to 2011 (Lon *et al.*, 2014; Lin *et al.*, 2015). Subjects found to have *P. vivax* malaria were treated with dihydroartemisinin-piperaquine (DP), then followed for recurrence with weekly blood smears for six weeks and with a monthly blood smear after that. Of 78 *P. vivax*-infected subjects followed for a median of 115 days, 23 individuals, or approximately one third of the cohort, developed a recurrent infection. These recurrences likely represent relapse or reinfection, because treatment failure with DP is unlikely. During the follow-up, six subjects suffered second recurrences, and one subject suffered a third recurrent infection. Hence, a total of 30 recurrent infections were available for the follow-up genotype analysis. In combination with 78 subjects at the baseline, there are 108 isolates available for the genotype analysis. To avoid the bias due to length of follow-up, we only use 78 baseline sequencings ($n = 78$) to estimate the parameters in relapse model (1), with 23 positive responses ($\sum_{i=1}^{n} Y_i = 23$). However, in order to utilize as much information as possible, we include those 7 second or third recurrences in the estimation of transition probabilities q and q^*, using their most recent sequencing as the baseline sequencing. This results in 30 pairs of baseline and recurrence sequencings in the log-likelihood function (7) with $m = 30$.

TABLE 3 Operating characteristics of proposed classifiers under low-dimensional scenarios

			BPM			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
Scenario	μ	n	**sens**	**spec**	**acc**	**sens**	**spec**	**acc**	**sens**	**spec**	**acc**
1	0.05	100	89.1	83.3	88.2	89.6	12.0	76.5	93.2	82.4	91.2
		200	89.2	83.7	88.3	97.0	3.8	81.8	98.4	87.2	96.5
		400	89.3	84.2	88.5	99.5	0.7	83.4	98.8	87.8	97.0
		800	89.4	84.0	88.5	100.0	0.0	83.7	98.9	88.3	97.2
	0.12	100	88.9	84.4	87.5	64.7	38.2	56.1	73.0	79.5	75.0
		200	89.1	84.2	87.5	78.4	25.3	60.9	93.2	91.2	92.5
		400	89.2	84.2	87.6	87.0	16.8	63.9	97.4	92.0	95.6
		800	89.3	84.1	87.6	92.2	11.5	65.7	97.6	92.3	95.9
	0.25	100	89.1	84.4	86.7	44.3	58.5	51.6	50.8	72.8	61.9
		200	89.4	84.3	86.8	46.1	57.6	52.0	71.1	90.1	80.8
		400	89.3	84.4	86.9	50.9	55.1	53.1	92.0	94.4	93.2
		800	89.4	84.3	86.8	53.0	55.3	54.2	95.7	94.7	95.2
2	0.05	100	89.2	85.1	88.3	81.4	19.8	68.5	85.8	79.5	84.3
		200	89.3	84.9	88.4	92.3	8.3	74.9	96.7	88.4	94.9
		400	89.4	84.5	88.4	98.2	2.0	78.5	98.4	88.8	96.4
		800	89.5	84.2	88.5	99.8	0.2	79.3	98.5	89.5	96.6
	0.12	100	89.1	84.5	87.2	55.3	48.0	52.4	61.9	74.4	66.6
		200	89.2	83.9	87.1	63.3	39.3	53.7	83.9	90.0	86.3
		400	89.5	83.9	87.3	73.8	29.7	56.4	96.2	93.0	94.9
		800	89.6	84.0	87.4	80.2	23.6	57.8	97.0	93.4	95.6
	0.25	100	89.3	84.1	86.3	41.1	62.5	53.6	44.6	70.5	59.5
		200	89.6	84.1	86.4	35.6	67.5	54.1	54.7	87.3	73.6
		400	89.8	84.2	86.5	35.9	67.7	54.3	80.8	94.6	88.7
		800	89.7	84.2	86.5	33.7	71.7	55.6	94.0	95.5	94.9

TABLE 4 Average bias of coefficient estimation, variable selection, and operating characteristics of our proposed classifier under high-dimensional scenarios

Scenario 3					$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
μ	(n, J)	**Bias**	**TP**	**TN**	**sens**	**spec**	**acc**	**sens**	**spec**	**acc**
0.05	(100,200)	0.11	5	146	100	0	92.0	100	80.6	98.3
	(200,400)	0.03	4	332	100	0	92.2	99.0	97.0	98.8
0.12	(100,200)	0.07	5	157	99.0	1.3	82.3	100	69.1	94.6
	(200,400)	0.02	5	353	99.8	0.1	83.2	99.0	85.1	96.7
0.25	(100,200)	0.09	5	156	89.1	11.9	65.9	100	75.2	92.6
	(200,400)	0.04	5	351	82.4	17.5	63.7	89.9	85.8	88.9
Scenario 4					$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
μ	(n, J)	**Bias**	**TP**	**TN**	**sens**	**spec**	**acc**	**sens**	**spec**	**acc**
0.05	(100,200)	0.01	5	194	100	0	91.9	100	50	95.2
	(200,400)	0.01	5	395	100	0	93.6	100	80	98.2
0.12	(100,200)	0.07	5	157	90.9	0	76.9	97.6	57.1	91.1
	(200,400)	0.02	5	354	97.4	0	83.5	97.7	88.9	96.7
0.25	(100,200)	0.05	5	174	70.5	27.5	59.2	86.7	74.3	80.9
	(200,400)	0.11	5	280	71.8	24	58.8	96.3	83.3	92

Targeted deep sequencing was performed on DNA extracted from filter paper blood spots collected by finger prick. A nested polymerase chain reaction (PCR) assay was used to amplify a 117-base pair variable portion of the *P. vivax* merozoite surface protein 1 (*pvmsp1*) gene based on previous work showing great nucleotide diversity across this region (Parobek *et al.*, 2014). Samples were amplified in duplicate and individually tagged, then pooled and sequenced on the Ion Torrent platform from Life Technologies. The *Pvmsp1* sequence variants were determined by SeekDeep, a bioinformatics pipeline that uses a clustering method to construct the most likely haplotypes within a patient while removing false haplotypes due to PCR or sequencing error (Hathaway *et al.*, 2018). For each subject, *pvmsp1* haplotypes that were present in two independent duplicate PCR samples at ≤0.5% frequency were counted as unique variants. Consensus haplotypes were each assigned a unique population identifier based on their prevalence in the cohort, namely, CAM.00 to CAM.66 with CAM.00 being the most prevalent *pvsmp1* variant encountered. In total, 67 unique *pvmsp1* variants were detected across 108 isolates. Nine common variants appeared in at least 10% of individuals, while two-thirds of variants appeared in only one isolate. In-host genetic diversity was also high, as 90% of initial infections contained multiple variants, displaying an average of 3.6 co-circulating variants.

We used the penalized likelihood model with an L_1-penalty, as shown in (3), with 5% reinfection rate. We report variants in the initial and recurrence sequencing, their estimated coefficients $\widehat{\boldsymbol{\beta}}$ in model (1), prevalence of the variants, two classification probability estimates, and classification results based on $\widehat{\xi}_i^{(1)}$ and BPM method. Variants with a nonzero estimated coefficient are considered to be associated with relapse. Using the profiled likelihood function (7), the maximum likelihood estimates for the transition probability are $\widehat{q} = 0.387$ and $\widehat{q}^* = 0.987$.

Table 5 shows part of the classification results. A complete list of the classification results is shown in the Supporting Information. First, one can see that the recurrence is likely classified as reinfection if variants in the recurrence are prevalent and not observed in the initial sequencing. Taking 151→151R pair, for example, the nonsharing variant CAM.00 that appeared in the recurrence sequencing is the most prevalent variant in the sample, suggesting that the recurrence is likely reinfection. Second, the high transition probability $\widehat{q}^* = 0.987$ suggests that an unobserved variant in the initial sequencing likely remains unobserved in the recurrence sequencing if the recurrence is a relapse. This explains why 152→152R pair is classified as reinfection. The appearance of prevalent variants CAM.05 and CAM.07 in the recurrence sequencing significantly lowers the classification probability from $\xi_i^{(0)}$ to $\xi_i^{(1)}$.

Some recurrence pairs tend to have more diverse and abundant minority variants. Many variants tend to be nonsharing due to this abundance. For example, both 80→80R pair and 125→125R pair have multiple variants in the recurrence sequencing that did not appear in the initial sequencing, resulting in a low value of $\xi_i^{(1)}$ and reinfection as the classification result. It is worth noting that, nonsharing variants in the initial sequencing have little impact on $\xi_i^{(1)}$. Taking 36→36R, for example, the pair is still classified as relapse even when seven initial variants were not observed in the recurrence sequencing. The classification probability $\xi_i^{(1)}$ only slightly decreases from $\xi_i^{(0)}$. This tendency can be explained by a low value of transition probability estimate $\widehat{q} = 0.387$. If $\widehat{q}$ is small, it is not unusual to see a variant in the initial sequencing not observed in the recurrence sequencing if the infection is a relapse pair like 36→36R. In contrast, even though the 80R→80RR pair has five sharing variants in the initial sequencing, the classification probability $\xi_i^{(1)}$ decreases significantly from $\xi_i^{(0)}$ because one nonsharing variant in the recurrence sequencing, CAM.04, is prevalent.

When comparing our method to the BPM, disparity occurs when prevalent variants appeared only in the recurrence sequencing. As discussed earlier, the 152→152R and 80R→80RR pairs are classified as reinfection by our method because nonsharing variants appeared in recurrence sequencing are prevalent. The BPM method classifies them as relapse because more than one prevalent variant overlapped in both sequencings, such as CAM.00 and CAM.01 in the 152→152R pair, and CAM.00, CAM.02, and CAM.06 in the 80R→80RR pair. Contrarily, the BPM method likely classifies a recurrent infection to reinfection if there is only one sharing variant that is prevalent, such as CAM.00 in the 96→96R pair. Our method otherwise classifies the pair as relapse because there are not enough nonsharing variants appeared in the recurrence sequencing. In summary, the classification result of 80→80R pair demonstrates the flaw of BPM. When multiple prevalent nonsharing variants (such as CAM.01, CAM.02, and CAM.03) appears in the recurrence, it is more likely the recurrence is reinfection, not relapse. A method like BPM considering only shared variants ignores this possibility and likely misclassifies the case.

Note that, from a statistical point of view, the analysis is sensitive to the selection of background reinfection rate. If the reinfection rate is misidentified, the maximum likelihood estimator of the coefficients in model (1) may not be consistent, as well as the classification probability $\widehat{\xi}_i^{(0)}$ calculated from these estimators. The classification probability $\widehat{\xi}_i^{(1)}$ may not be consistent as well because it is established on the initial classification probability $\widehat{\xi}_i^{(0)}$. In this data analysis, the classification result based on $\widehat{\xi}_i^{(1)}$ is quite robust when the reinfection rates is less than 10%. Meanwhile, an in vivo study on the dynamics of *P. vivax* infection suggests that up to 96% of the *P. vivax* infection is due to relapse in individuals living in the endemic areas in Thailand (Adekunle *et al.*, 2015). Cambodia is in Southeast

TABLE 5 Classification results for recurrence pairs using our method and binomial probability model

Recurrence Pair	Initial Variants	$\hat{\beta}$	$\hat{\xi}_i^{(0)}$	Recurrence Variants	Variant Prevalence	$\hat{\xi}_i^{(1)}$	Proposed Class	BPM Class
36→ 36R	CAM.00	1.833	0.960	CAM.01	0.269	0.870	Relapse	Relapse
	CAM.01	0.469		CAM.02	0.410			
	CAM.02	0.892		CAM.07	0.192			
	CAM.03	0		CAM.17	0.064			
	CAM.04	3.519						
	CAM.05	−1.085						
	CAM.06	−1.416						
	CAM.07	1.750						
	CAM.09	0						
	CAM.11	0						
80 → 80R	CAM.00	1.833	0.992	CAM.00	0.590	0.000	Reinfection	Relapse
	CAM.04	3.519		CAM.01	0.269			
	CAM.05	−1.085		CAM.02	0.410			
	CAM.08	0.395		CAM.03	0.295			
	CAM.09	0		CAM.05	0.231			
	CAM.24	2.954		CAM.06	0.231			
	CAM.27	0		CAM.07	0.192			
				CAM.08	0.154			
				CAM.12	0.064			
				CAM.41	0.013			
80R→ 80RR	CAM.00	1.833	0.673	CAM.00	0.590	0.340	Reinfection	Relapse
	CAM.01	0.469		CAM.02	0.410			
	CAM.02	0.892		CAM.04	0.346			
	CAM.03	0		CAM.06	0.231			
	CAM.05	−1.085		CAM.08	0.154			
	CAM.06	−1.416		CAM.12	0.064			
	CAM.07	1.750		CAM.59	0.013			
	CAM.08	0.395						
	CAM.12	0.677						
	CAM.41	0						
96 → 96R	CAM.00	1.833	0.979	CAM.00	0.590	0.992	Relapse	Reinfection
	CAM.02	0.892		CAM.30	0.013			
	CAM.04	3.519						
	CAM.08	0.395						
125 → 125R	CAM.02	0.892	0.130	CAM.00	0.590	0.000	Reinfection	Reinfection
				CAM.01	0.269			
				CAM.02	0.410			
				CAM.04	0.346			
				CAM.09	0.077			
				CAM.13	0.013			
				CAM.14	0.026			
				CAM.38	0.013			
				CAM.45	0.013			

(Continues)

TABLE 5 (Continued)

Recurrence Pair	Initial Variants	$\widehat{\beta}$	$\widehat{\xi}_i^{(0)}$	Recurrence Variants	Variant Prevalence	$\widehat{\xi}_i^{(1)}$	Proposed Class	BPM Class
151 → 151R	CAM.03	0	0.030	CAM.00	0.590	0.005	Reinfection	Reinfection
	CAM.05	−1.085		CAM.08	0.154			
	CAM.08	0.395		CAM.14	0.026			
152 → 152R	CAM.00	1.833	0.379	CAM.00	0.590	0.018	Reinfection	Relapse
	CAM.01	0.469		CAM.01	0.269			
				CAM.05	0.231			
				CAM.07	0.192			

Asia and geographically adjacent to Thailand. Assuming 5% reinfection rate in this area is reasonable. Interestingly, from the complete list of our classification result in the Supporting Information, 23 individuals had recurrent infections among 78 subjects at the baseline. Among those 23 subjects, 10 subjects are classified as reinfections by our algorithm. The reinfection rate $\mu = P(N_i = 1|R_i = 0)$ may be estimated at 10/(78-13) = 15%, which is higher than the literature suggests.

6 | DISCUSSION

In this paper, we propose a novel classification method that is model-based and utilizes transition likelihoods to classify recurrent *P. vivax* infections as either relapse or reinfection. Previous work used only shared variants to calculate the reinfection probability. Here, we show that nonshared variants are also informative. Both simulation studies and real data analysis support the feasibility and practical use of our classifier. Some assumptions and generalizations of our method are worth of discussion.

First, we assume that the reinfection rate μ is known or can be correctly specified. Model misspecification on μ can be problematic for both regression coefficient estimation and classification probability calculation when an incorrect value is used. Through simulation experiments listed in our Supporting Information, one can see the impact of the misspecification is apparent when the sample size is small. When the sample size is large, however, the bias in the coefficient estimation diminishes, and the performance of our classifier improves. Our approach is robust to the misspecification of μ when the sample size is large. As one can imagine, bias more likely occurs to the estimation of intercept α because both μ and α represent some sense of baseline occurrence rates. Underestimation of μ shall lead to overestimation of α, and overestimation of μ shall lead to underestimation of α to balance the overall baseline occurrence rate. Such tendency in bias can be seen in our simulation results in the Supporting Information. Meanwhile, although misspecification on μ leads to biased estimation of $\xi_i^{(0)}$, the classification performance of $\widehat{\xi}_i^{(1)}$ utilizing transition likelihoods is mildly affected when the reinfection rate is underestimated. Even when the reinfection rate is overestimated, the accuracy of the classifier $I(\widehat{\xi}_i^{(1)} > 0.5)$ can still reach a satisfactory level.

Second, we assume the occurrence of the variants is independent. This assumption can be checked in our real data. Using Fisher's exact tests for presence/absence of any two of 13 most frequent variants, the minimum p-value is 0.0048 and only 10 out 78 pairs have P-value smaller than 0.05. After Benjamini-Hochberg adjustment for multiple comparisons, none of the P-values is smaller than 0.05. The independence assumption is not significantly violated in our case.

Finally, we assume the transition probabilities are equal for all variants, that is, $q_1 = \cdots = q_J$ and $q_1^* = \cdots = q_J^*$. This assumption can be relaxed using external information to model the transition probabilities. Specifically, one can assume the probability follow a logistic model $\log\{q_j/(1 - q_j)\} = W_j'\gamma$ and $\log\{q_j^*/(1 - q_j^*)\} = W_j'\gamma^*$, where W_j is a column vector of external covariates, and γ and γ^* are column vectors of regression coefficients. In our case, reading frequency of the variant may be the covariate that is associated with the transition of the variants. It is worth noting that we assume the transition starts from a new infection to either relapse or reinfection. However, in our real data, there are seven second or third recurrent infections. Although in the real data analysis we treated the most recent infection as the initial infection, the recurrent infection may depend on multiple previous events in this case. The modeling is much more complicated, considering the status of previous infections is unknown except for the baseline infection. It is not clear whether a relapse infection could be associated with the transition probability of the variants. One possible approach is to include the relapse indicator, R_i, in the logistic model for q_j and q_j^*, as part of covariates W_j. However, because R_i is not observable, it is not clear how γ and γ^* can be estimated. We leave it for future research.

Our current analysis considers only recurrence indicator without time domain involved. The causes of the recurrent infection can actually be seen as competing risks, for which one observes the event occurrence of either relapse or

reinfection. In our case, the cause of the event is unknown in all of the events, so the challenge remains as how one can derive the classification probability using a hazard model and transition likelihoods to classify the recurrent infections incorporating the time to infection.

ACKNOWLEDGMENTS

The authors wish to thank editor, associate editor, and reviewers for their careful review and constructive comments. This work is partially supported by National Institutes of Health, through grant award number UL1TR002489 and K08AI110651.

DATA AVAILABILITY STATEMENT

The malaria infection data used in the real data analysis are also included in the supporting information and available for public use.

ORCID

Feng-Chang Lin https://orcid.org/0000-0002-2638-1775
Quefeng Li https://orcid.org/0000-0003-0707-2763
Jessica T. Lin https://orcid.org/0000-0002-4516-723X

REFERENCES

Adekunle, A.I., Pinkevych, M., McGready, R., Luxemburger, C., White, L.J., Nosten, F., Deborah, C. and Davenport, M.P. (2015) Modeling the dynamics of *Plasmodium vivax* infection and hypnozoite reactivation *in vivo*. *PLoS Neglected Tropical Diseases*, 9, e0003595.

Beck, H.-P., Wampfler, R., Carter, N., Koh, G., Osorio, L., Rueangweerayut, R., Krudsood, S., Lacerda, M.V., Llanos-Cuentas, A., Duparc, S., Rubio, J.P. and Green, J.A. (2016) Estimation of the antirelapse efficacy of tafenoquine, using plasmodium vivax genotyping. *The Journal of Infectious Diseases*, 213, 794–799.

Chen, N., Auliff, A.M., Rieckmann, K.H., Gatton, M.L. and Cheng, Q. (2007) Relapses of *Plasmodium vivax* infection result from clonal hypnozoites activated at predetermined intervals. *The Journal of Infectious Diseases*, 195 (7), 934–41.

Daniels, R., Volkman, S.K., Milner, D.A., Mahesh, N., Neafsey, D.E., Park, D.J., Rosen, D., Angelino, E., Sabeti, P.C., Wirth, D.F. and Wiegand, R.C. (2008) A general SNP-based molecular barcode for *Plasmodium falciparum* identification and tracking. *Malaria Journal*, 7, 223.

Fan, J. and Lv, J. (2011) Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57, 5467–5484.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

Hathaway, N.J., Parobek, C.M., Juliano, J.J. and Bailey, J.A. (2018) SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Research*, 46, e21.

Imwong, M., Snounou, G., Pukrittayakamee, S., Tanomsing, N., Kim, J.R., Nandy, A., Guthmann, J.P., Nosten, F., Carlton, J., Looareesuwan, S., Nair, S., Sudimack, D., Day, N.P., Anderson, T.J. and White, N.J. (2007) Relapses of *Plasmodium vivax* infection usually result from activation of heterologous hypnozoites. *The Journal of Infectious Diseases*, 195, 927–933.

Juliano, J.J., Ariey, F., Sem, R., Tangpukdee, N., Krudsood, S., Olson, C. LooareesuwanS. Rogers, W.O., Wongsrichanalai, C. and Meshnick S.R., (2009) Misclassification of drug failure in *Plasmodium falciparum* clinical trials in Southeast Asia. *The Journal of Infectious Diseases*, 200, 624–628.

Juliano, J.J., Porter, K., Mwapasa, V., Sem, R., Rogers, W.O., Ariey, F., Wongsrichanalai, C., Read, A. and Meshnick, S.R. (2010) Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proceedings of the National Academy of Sciences*, 107, 20138–20143.

Kobbe, R., Neuhoff, R., Marks, F., Adjei, S., Langefeld, I., von Reden, C. AdjeiO. Meyer, C.G. and May, J. (2006) Seasonal variation and high multiplicity of first *Plasmodium falciparum* infections in children from a holoendemic area in Ghana, West Africa. *Tropical Medicine & International Health*, 11, 613–619.

Kwiek, J.J., Alker, A.P., Wenink, E.C., Chaponda, M., Kalilani, L.V. and Meshnick, S.R. (2007) Estimating true antimalarial efficacy by heteroduplex tracking assay in patients with complex *Plasmodium falciparum* infections. *Antimicrobial Agents and Chemotherapy*, 51, 521–527.

Lin, J.T., Hathaway, N.J., Saunders, D.L., Lon, C., Balasubramanian, S., Kharabora, O. GosiP. Sriwichai, S., Kartchner, L., Chuor, C.M., Satharath, P., Lanteri, C., Bailey, J.A. and Juliano, J.J. (2015) Using amplicon deep sequencing to detect genetic signatures of *Plasmodium vivax* relapse. *The Journal of Infectious Diseases*, 212, 999–1008.

Lon, C., Manning, J.E., Vanachayangkul, P., So, M., Sea, D., Se, Y., Gosi, P., Lanteri, C., Chaorattanakawee, S., Sriwichai, S., Soklyda, C., Kuntawunginn, W., Buathong, N., Nou, S., Walsh, D.S., Tyner, S.D., Juliano, J.J., Lin, J., Spring, M., Bethell, D., Kaewkungwal, J., Tang, D., Chuor, C.M., Satharath, P. and Saunders, D. (2014) 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*, 9, 1–13.

Nyachieo, A., Van Overmeir, C., Laurent, T., Dujardin, J.-C. and D'Alessandro, U. (2005) *Plasmodium falciparum* genotyping by microsatellites as a method to distinguish between recrudescent and new infections. *The American Journal of Tropical Medicine and Hygiene*, 73, 210–213.

Parobek, C.M., Bailey, J.A., Hathaway, N.J., Socheat, D., Rogers, W.O. and Juliano, J.J. (2014) Differing patterns of selection and geospatial genetic diversity within two leading *Plasmodium vivax* candidate vaccine antigens. *PLoS Neglected Tropical Diseases*, 8, 1–17.

Pearson, R.D., Amato, R., Auburn, S., Miotto, O., Almagro-Garcia, J., Amaratunga, C. SuonS. Mao, S., Noviyanti, R., Trimarsanto, H., Marfurt, J., Anstey, N.M., William, T., Boni, M.F., Dolecek, C., Hien, T.T., White, N.J., Michon, P., Siba, P., Tavul, L., Harrison, G., Barry, A., Mueller, I., Ferreira, M.U., Karunaweera, N., Randrianarivelojosia, M., Gao, Q., Hubbart, C., Hart, L., Jeffery, B., Drury, E., Mead, D., Kekre, M., Campino, S., Manske, M., Cornelius, V.J., MacInnis, B., Rockett, K.A., Miles, A., Rayner, J.C., Fairhurst, R.M., Nosten, F., Price, R.N. and Kwiatkowski, D.P. (2016) Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. *Nature Genetics*, 48, 959–964C.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. The program codes for the real data analysis are also available at the website.

How to cite this article: Lin F-C, Li Q, Lin JT. Relapse or reinfection: Classification of malaria infection using transition likelihoods. *Biometrics*. 2020;1–13. https://doi.org/10.1111/biom.13226